Psychometrika

VOLUME VII—1942
JANUARY-DECEMBER

PUBLISHED QUARTERLY
By]THE PSYCHOMETRIC SOCIETY
AT 23 WEST COLORADO AVENUE
COLORADO SPRINGS, COLORADO

Factor

Psychometrika

CONTENTS

VOLUME SEVEN MARCH 1942 NUMBER ONE

FREQUENCY OF PUBLICATION

Even though PSYCHOMETRIKA faces a possible increase in publication cost and a reduction in revenue, due to loss of many foreign and some domestic subscribers, authors will continue to receive 200 reprints of their articles free of charge. Furthermore , publication charges to authors which were suspended in 1941 will not be resumed in 1942. However, in order to maintain Psychometrika on a sound financial basis, it has been decided to revert to quarterly publication beginning with March, 1942 and continuing until general conditions warrant more frequent publication.

PSYCHOMETRIKA—VOL. 7, NO. 1
MARCH, 1942

THE PSYCHOMETRIC SOCIETY—ROOTS AND POWERS*

JACK W. DUNLAP
UNIVERSITY OF ROCHESTER

A classical function familiar to all psychologists is the expression $S \rightarrow R$, commonly called the $S-R$ bond, or stimulus-response. This function might well be the motto of this Society since one of the primary concerns of the Society's members is that of prediction and control; that is, given a particular stimulus or set of stimuli, what is the most likely expectation as to the response or result. These symbols are particularly significant when one examines them in connection with the Society, for it leads one to a consideration of the causes that led to its organization and to its probable effect on psychology in the future.

The Gestaltist might interpret the formation of the Society as an illustration of the principle of closure, but, while this is a perfectly reasonable rationalization of the *fait accompli,* it seems to have a tone of finality and completeness which is not at all in keeping with the possibilities of the future for the Psychometric Society. For this reason, I have chosen the Thorndikian rather than the Gestalt principle to exemplify our Society.

Psychology as a formal study is young as disciplines go, and, as a science is barely through its birth pains. In considering this science one cannot but remember James' famous remark about the new-born child being in a state of "blooming, buzzing confusion" and realizing that psychology is just beginning to attain some order in its domain. This state of confusion is not peculiar to psychology or to psychologists but is natural in any new field of human endeavor. The original approaches in any science are based on common experiences, philosophical speculations, and inherited superstitions, and only with exasperating slowness can a body of quantitative data based on controlled observations be secured from which it is possible to develop rational hypotheses.

The task confronting the early workers in the field of psychology was tremendous, and there is no obvious and easy way for us to evaluate the contributions to quantitative rational psychology of such men

* Presidential Address, Psychometric Society, September 4, 1941, Northwestern University.

as Bacon and Galton in experimental design, of Weber and Fechner in psychophysics, of Lashley and Rashevsky in physiological psychology, of Binet, Terman, and Thorndike in mental measurements, of Spearman and Thomson in quantifying theories of intelligence, of Allport in social psychology, of Warden in formulating the dynamics of animal behavior, and of Fisher, Kelley, Pearson, and Thurstone in the analysis of data, to mention only a few. The efforts of these men are not only widely spread over the field of psychology, but they have been even more widely scattered in terms of time and geography. This long and widespread attempt at quantification is indicative of the need for such work. The importance of such attempts is further evidenced by the steadily increasing volume of technical literature appearing in various periodicals during the first four decades of the current century.

There can be no question that the stimuli were present, and only a catalyst was needed in the form of some individual to implement the uniting of widely separated scholars into an articulate and functional organization. In 1931 such an individual appeared: Dr. A. P. Horst. Horst had a firm conviction that there was a strong and growing interest in the quantification of our science and that what was needed was a medium of publication devoted to this purpose. He believed that the quantity and quality of the articles then appearing in widely scattered sources furnished a sound basis for establishing a journal devoted to the development of psychology on a quantitative rational basis.

In 1931 Horst was attempting to develop or locate a journal that would be devoted to quantitative methods as applied to education and psychology. The journals that most nearly met this condition were the *Journal of Educational Psychology* and the *Journal of the American Statistical Association.* Both of these, however, had other and more general purposes to serve than that proposed by Horst. During the following years Horst discussed the matter at great length with A. K. Kurtz and in 1933 carefully examined the possibilities of such a journal with L. L. Thurstone and M. W. Richardson. The idea of such a journal appealed strongly to Thurstone, since he was just beginning to publish his results on factor analysis. Richardson's interest in the theoretical problems of test construction guaranteed his support of the projected publication. During the latter part of 1933 the matter was brought to the attention of the speaker, because of his connection with the *Journal of Educational Psychology* and his consequent knowledge of the quantity of technical material available for the support of such a journal. In the spring of 1934 Thurstone went over the details of establishing such a journal, and at this time various methods of financing the journal were considered. Several attempts were made to interest one of the Foundations in supporting the proposed periodical, but all to no avail. Throughout the spring and summer of 1934 Horst, Kurtz, Richardson, and Stalnaker were working on details as to costs, publishers, policies, and the methodology of editorial management.

Thus, at the time of the fall meeting of the American Psychological Association at Columbia University, "Psychometrika" was only a nebula in a mist of words and wishes. A series of conferences of those interested in the project during the week of the meetings crystallized the plans, brought the group together as a unit, with the result that *Psychometrika* began to assume form and substance. As a result of these conferences the material gathered by different individuals as to publishers, costs, sales, style, policy, and editorial management were collated, and specific tasks were assigned to particular individuals. It was at this time that Kurtz began to emphasize the fact that, if there were readers for such a journal, they would be interested, in all likelihood, in forming a society in which their common interest would be the keynote.

The formation of such a society would have many advantages—the identifying of individuals with common interests, focusing attention on the need and importance of developing a quantitative rational psychology, providing a physical meeting where technical papers could be read (and perhaps, appreciated), of locating possible contributors to the journal, and last but not least, if the journal was to be the official organ of the society, providing financial support for its publication. The only fly in the ointment was that there was no idea of how many individuals were interested in such an organization. There seemed to be a number of cogent arguments to the effect that such an organization would have a greater chance of success if the journal, *Psychometrika,* were to appear, like Minerva from Jove's forehead, full blown before the public immediately after the organization of the society. But here a paradoxical situation arose—to have the journal it was first necessary to have the society, but to have the society it was claimed that one must first have the journal. So the matter stood through the fall of 1934 and the spring of 1935.

The next problem was to determine whether other biometricians, educators, psychologists, and statisticians were interested in forming such a Society. Thurstone made this possible by the liberal contribution of not only his own time and effort but also that of his staff. Through the facilities at his command, letters of inquiry were sent to a large number of individuals who, it was thought, might be interested. As a result of this canvass, invitations were extended to all who replied to attend the formation of the Society on September 4, 1935, at Ann Arbor, Michigan, during the session of the American Psychological Association. Temporary officers were appointed for the Society, and later in the fall a mail ballot for election of officers was held. Dr. L. L. Thurstone was the first president, Dr. Paul Horst, the secretary, and the speaker, the treasurer.

During the following year the constitution committee, composed of Horst, Kurtz, and Richardson, prepared the present constitution of the Society, which was officially adopted at the second annual meeting held at Dartmouth. This was amended in 1937 to include a "student membership," and at present such memberships constitute approximately one-sixth of the total membership.

The growth of the organization has been slow, but, on the other hand, the membership has had relatively few withdrawals. The paid membership for 1936 included 133 individuals and with the succeeding years included 185, 189, 214, and in 1940 dropped to 200 paid members. That the membership takes an interest in the affairs of the Society is indicated by the fact that approximately forty per cent of the eligible members voted in the recent elections.

In dealing with the historical development of the Society it is impossible to disentangle its history from that of the Psychometric Corporation. As pointed out above, one fundamental question was how to publish the Journal immediately upon the organization of the Society. At the Ann Arbor meeting it was voted to have dues of one dollar a year until the Journal appeared, and thereafter of five dollars a year. Thus, there was still no capital for starting the journal.

Suddenly shortly after the beginning of 1936, Horst became impatient, and with a confidence equalling his foresight, he cut the Gordian knot by offering to underwrite the losses of the journal for the first year up to one-fourth of its cost. A simple but practical solution, and an example which was immediately followed to a lesser extent by Kurtz, Thurstone, Richardson, and, so as not to appear too niggardly, by Dunlap. Somehow the word got about as to the plans for the journal and how it was to be financed initially. It was only a short time until pledges of support had been received from Guilford, Gulliksen, Kuder, Lorge, Stalnaker, and Thorndike. Suddenly, it was realized that sufficient money had been guaranteed to publish the journal for at least a year and a half, but offers still came in to help underwrite the venture. This spontaneous reaction seemed to more than justify the attempt to organize the Psychometric Society and to proceed with the publication of *Psychometrika.*

If the journal were to appear immediately, it was necessary that some legal unit be responsible for the financial arrangements. This was the basic reason for the formation of the Psychometric Corporation, independent of the Psychometric Society. Another cogent reason was that the original sponsors felt that during the early years the policy of the journal should conform closely with the basic ideas of its founders and not degenerate into another periodical devoted to publishing only the results of psychological and educational measurements. It was believed that if an editorial policy was firmly established, the journal would then go on regardless of changes in the composition of the editorial board. If, as time went on, the journal was a success, the Corporation would gradually turn over the control of *Psychometrika* to the members of the Society. It was for these reasons that on August 24, 1936, the Psychometric Corporation was incorporated in the State of Illinois.

That there was a real need for such a journal is shown by the list of library subscriptions, which has grown to include 78 libraries at present, exclusive of those in foreign countries that have dropped their subscriptions for the duration of the war. Within the short space of two years after its first appearance *Psychometrika* could be found in libraries in Canada, China, England, Austria, France, Germany, Scotland, South Africa, and a number of countries which now are only memories. This growth occurred in spite of the fact that the cost to a library was twice that for a private subscription. An unusual innovation was sending to libraries a fresh volume of the journal at the end of each year for binding purposes.

Originally the journal assessed contributors a dollar a page for text and tables and for the cost of cuts, but at the Columbus meeting in 1938, this charge was reduced to fifty cents a page with no charge for cuts. However, at the annual meeting held at Penn State College in 1940 this charge also was completely eliminated. The contributor has always been furnished *gratis* with two hundred copies of his article.

Another interesting fact about the journal is that each manuscript is first examined by the Managing Editor who removes all references to the name of the author. The manuscript is then evaluated by three members of the Editorial Board. The Managing Editor collates the comments on the manuscript and then accepts the article; rejects the article; or accepts it, conditional upon revisions suggested by the readers. This practice has contributed in no small way to the quality and uniform grade of material appearing in the Journal. The untiring efforts of the Editor, Dr. M. W. Richardson and of Dr. Dorothy Adkins, Assistant Editor, have been no small factor in the production of the Journal.

Publication of a journal is not all that the Corporation has done.

In 1938 there appeared the first of a series of *Psychometric Monographs.* This first monograph, entitled "Primary Mental Abilities," illustrates with emphasis the type of quantitative rational psychology that the Society has tried to develop. In 1939 the Psychometric Corporation made possible the publication of another journal, the *Bulletin of Mathematical Biophysics.* The Corporation acted as publishing and financial agent for the *Bulletin* and in 1940 turned this publication over to the University of Chicago Press.

Today we are a society on a firm professional and financial basis, and we have a journal that has attained a high rank among scientific periodicals. It is with no little pleasure that, as a former treasurer of both the Society and the Corporation, I announce that the loans of the original sponsors have been repaid and that both the Society and Corporation have no debts but rather a small but comfortable reserve with which to face the contingencies of the future.

But enough of the past. Let us consider the present and the future of our Society. At each annual meeting the Society has sponsored a program of papers. Last year barely enough papers were submitted to form a program, and this year the number of papers submitted was so small that it was impossible to arrange a program. What does this mean? Is it that our members are not engaged in productive scientific work? Is interest in the development of the principles of the organization waning? Is the type of program not satisfactory? Are our members so busy with affairs of national defense that they cannot participate in scientific programs?

I do not know, but I suspect that not one but all of these reasons have contributed to a greater or lesser extent to this unfortunate state of affairs. This is an important juncture in the development of the Psychometric Society. Shall we abandon our program of papers? Shall we have a program consisting of two or three invited papers? Or shall we attempt to design an entirely new type of program? This is not a task for the President or for that matter, for your officers, but this is a task for the total membership. Surely the time has not arrived when we can comfortably recline on our advances and say, "Psychology is now a quantitative rational science; let us maintain the status quo." That there is work to be done, almost an unlimited amount, seems to be apparent. What must be done is to induce each member to take a more active part in the affairs of the Society and to contribute of his time, energy, and ideas. Your officers will welcome any suggestions and be glad to receive your assistance in implementing these suggestions into concrete action.

Perhaps part of the fault lies with our Journal, which has a preponderance of material so highly technical that only the specialist can read and understand it. But we, as members, must remember that the editors cannot publish material that is not submitted to them. Personally, I would like to see more material such as the highly practical article of Richardson and Kuder on the "Theory of Estimating Test Reliability." If the members want articles other than those on factor analysis or the determination of the order of a matrix, they must submit technical manuscripts on other topics.

The future should see a great many papers of theoretical and practical importance in all fields of psychology. That problems abound that demand formulation in quantitative and mathematical terms you know even better than I. The entire field of quotients needs to be examined and stated in more precise and rigorous terms. We have allowed ourselves to be shackled by the I.Q. with the resulting controversies about the stability of this function. Millions of words have been written on the subject, and thousands of hours have been expended in computing and recomputing this function. Curiously, little has been done on the basic rationale underlying the function. Thurstone and Thorndike have attempted to develop more satisfactory units of measurement, and Heinis has developed a function which has received far too little consideration in the discussions of this problem. Here indeed, is a field worthy of investigation and restatement.

The field of item analysis and test construction is just beginning to emerge from the labored gropings as to methodology. Recently it was my good fortune to examine a manuscript by Horst *et al,* in which a systematic attempt had been made to develop item analysis on a rational basis. This work, however, was far from conclusive, and I am sure the authors would be the first to disclaim that all the problems were solved.

The general field of prediction, which is being emphasized so strongly in this time of national emergency, when it is vital to our country and to each of us that each man be placed where he will be most effective, is filled with problems demanding our attention. What are criteria? How can their validity be established? What is the most predictable criterion? What is the minimum number of variables from a given matrix which will give valid, reliable, and effective estimation of either a single or a multiple criterion?

The rationale of rating scales has advanced little since the last world war, and there is no question but that the pressure of time and numbers will again bring such scales to the fore. Here, indeed, is a field that should challenge the membership, not only for its theoretical implications but also for its practical applications.

The current personality scales represent a cut-and-try methodology, and only recently has there been any attempt to apply modern methods of analysis to such scales. It is necessary in this field not only to apply more rigorous statistical techniques but also to attempt to delimit the problems more precisely and to formulate hypotheses susceptible to experimental study.

Whichever way one turns he is confronted with problems in social psychology, animal psychology, and in the psychology of personality, to mention a few, where solutions await the development of a rationale that can be subjected to quantitative formulation. I could go on with these citations, but why should I mention other fields when you are even more familiar with their problems than I?

The Psychometric Society emerged as a result of a felt need and so far has served its purpose admirably. That its services to psychology will be more substantial in the future is not merely a possibility, but is a probability of a very high order. The roots of the Society are firmly fixed, and its powers, though latent, are just beginning to emerge and will, I am confident, be a major force in the psychology of tomorrow.

PSYCHOMETRIKA—VOL. 7, NO. 1
MARCH, 1942

ON THE NUMBER OF FACTORS

QUINN MCNEMAR
STANFORD UNIVERSITY

A proposed criterion for the number of factors is developed on the basis of the similarity between a factorial residual and the partial correlation coefficient; something is known concerning the sampling error of the latter. Instead of computing the residuals as partials, a formula is presented for adjusting the standard deviation of the distribution of residuals so as to approximate the S.D. of the residuals as partial correlations. The criterion requires that factors be extracted until the adjusted S.D. reaches or falls below $1/\sqrt{N}$. When tried out on six samples drawn from six universes of known factorial description, the criterion indicated the correct number of factors each time. The requisites of situations adequate for such empirical checks are discussed.

It is appropriate to begin this paper with a few words regarding *a priori* requisites for an adequate criterion for the number of factors. It would not seem unreasonable to require that any proposed criterion should exhibit some degree of rationality before being tried out. Some of the proposals are rather obviously nonsensical, and therefore unworthy of consideration. Let us list a few of these and point out, rather categorically, some objections thereto.

First: the frequency distribution of centroid loadings should become uni-modal. This can never be adequate because the centroid method precludes uni-modality.

Second: the range of the centroid loadings should be less than some arbitrary value, *e.g.*, .30. This gets us nowhere since we would then need a criterion for setting up the arbitrary value.

Third: the curve of some defined function should flatten markedly. To be useful, such a criterion would need to be fortified with a criterion for deciding when a curve has flattened markedly.

Let us now turn to a few positive suggestions. It seems logical to suppose that an adequate criterion will be some function of the size of the sample. It would be strange indeed to find a formula concerned with sampling errors which is not a function of capital N. By analogy with the standard error of a distribution of tetrads one might expect an adequate criterion to be a function of the number of variables. On the practical side, a satisfactory criterion must not involve an unreasonable amount of computation. One analytically derived criterion calls for the value of the determinant of the original correlation matrix. We can foresee some difficulty in computing this for a determinant of order 57.

The first criterion ever proposed was that the analysis should be continued until the standard deviation of the residuals becomes less than the sampling error of the mean original correlation coefficient. This criterion has not proved valid in experimental studies, and it does not always work when checked empirically; but because of its credibility, we have re-examined it in order to see whether some modification might make it acceptable. Since the residuals are highly analogous to partial r's, one wonders why it should have been assumed that their sampling errors should be a function of the magnitude of the original correlations. This isn't the case for partial r's, so we suggest that it would be more logical to require the residual standard deviation to fall below the standard error of an r of zero rather than that for the mean original r. This would lead to the extraction of fewer factors, and thus would tend to exaggerate the known bias in the residual criterion.

But let us shove the analogy with the partial correlation technique a bit farther. In computing, say, the first factor residuals, the numerator term of the formula for partial r is being used, whereas the denominator terms are ignored. If these latter terms were used, the residuals would be larger, and they would correspond to partial r's. Now the standard error of a partial r is known, hence the significance of the deviation of such residuals from zero could readily be determined. Instead of disrupting the centroid process by computing residuals as partials, it is possible to adjust the standard deviation of the ordinarily obtained residuals so as to approximate closely the standard deviation of the distribution of the corresponding partials. Then the analysis would be carried to the point at which such an adjusted standard deviation is equal to or less than the standard error of a zero correlation, *i.e.*, less than $1/\sqrt{N}$.

Derivation

Let r_{ab} be the correlation between tests a and b, and let a_1, a_2, $a_3, \cdots, b_1$, b_2, $b_3, \cdots$ stand for centroid factor loadings. Suppose one factor has been extracted, the residual is

$$\rho_{ab} = r_{ab} - a_1 b_1 . \qquad (1)$$

If, however, this residual were computed as a partial correlation we would have

$$r_{ab\cdot 1} = \frac{r_{ab} - a_1 b_1}{\sqrt{1 - a^2{}_1}\,\sqrt{1 - b^2{}_1}}\,. \tag{2}$$

Now if a second factor has been extracted, the residual expressed as a partial correlation would be of the form

$$r_{ab\cdot 12} = \frac{r_{ab\cdot 1} - r_{a2\cdot 1}\, r_{b2\cdot 1}}{\sqrt{1 - r^2{}_{a2\cdot 1}}\,\sqrt{1 - r^2{}_{b2\cdot 1}}}\,. \tag{3}$$

But

$$r_{a2\cdot 1} = \frac{r_{a2} - r_{a1} r_{12}}{\sqrt{1 - r^2{}_{a1}}\,\sqrt{1 - r^2{}_{12}}}\,.$$

Since $r_{12} = 0$, $r_{a2} = a_2$, and $r_{a1} = a_1$, we have

$$r_{a2\cdot 1} = \frac{a_2}{\sqrt{1 - a^2{}_1}}$$

and

$$r_{b2\cdot 1} = \frac{b_2}{\sqrt{1 - b^2{}_1}}$$

Then (3) becomes

$$r_{ab\cdot 12} = \frac{\dfrac{r_{ab} - a_1 b_1}{\sqrt{1 - a^2{}_1}\,\sqrt{1 - b^2{}_1}} - \dfrac{a_2}{\sqrt{1 - a^2{}_1}} \cdot \dfrac{b_2}{\sqrt{1 - b^2{}_2}}}{\sqrt{1 - \dfrac{a^2{}_2}{1 - a^2{}_1}}\sqrt{1 - \dfrac{b^2{}_2}{1 - b^2{}_1}}}\,,$$

which simplifies to

$$r_{ab\cdot 12} = \frac{r_{ab} - a_1 b_1 - a_2 b_2}{\sqrt{1 - a^2{}_1 - a^2{}_2}\,\sqrt{1 - b^2{}_1 - b^2{}_2}}\,. \tag{4}$$

By a similar use of partials of lower order, it can be shown that the third factor residuals as partials take on the form

$$r_{ab\cdot 123} = \frac{r_{ab} - a_1 b_1 - a_2 b_2 - a_3 b_3}{\sqrt{1 - a^2{}_1 - a^2{}_2 - a^2{}_3}\,\sqrt{1 - b^2{}_1 - b^2{}_2 - b^2{}_3}}\,.$$

Presumably, with laborious algebra, this could be extended. But the use of determinants greatly facilitates the resolution for the general case of *s* factors. Let the major determinant of the system of any two tests and *s* factors be designated as

$$D = \begin{vmatrix} 1 & r_{ab} & a_1 & a_2 & a_3 & \cdots & a_s \\ r_{ab} & 1 & b_1 & b_2 & b_3 & \cdots & b_s \\ a_1 & b_1 & 1 & 0 & 0 & \cdots & 0 \\ a_2 & b_2 & 0 & 1 & 0 & \cdots & 0 \\ a_3 & b_3 & 0 & 0 & 1 & \cdots & 0 \\ . & . & . & . & . & \cdots & . \\ . & . & . & . & . & \cdots & . \\ . & . & . & . & . & \cdots & . \\ a_s & b_s & 0 & 0 & 0 & \cdots & 1 \end{vmatrix}$$

Then, by formula 266 of Kelley (4, p. 298)

$$r_{ab\cdot 123\cdots s} = \frac{D_{ab}}{\sqrt{D_{aa}}\,\sqrt{D_{bb}}}, \tag{5}$$

where the subscripts indicate which row and column have been deleted in defining the minors. Straightforward evaluation of these three minors lead to

$$r_{ab\cdot 123\cdots s} = \frac{r_{ab} - a_1 b_1 - a_2 b_2 - a_3 b_3 - \cdots - a_s b_s}{\sqrt{1 - a^2_1 - a^2_2 - a^2_3 - \cdots - a^2_s}\,\sqrt{1 - b^2_1 - b^2_2 - b^2_3 - \cdots - b^2_s}} .$$

It will be noted that the numerator is the ordinary residual and that the denominator is the product of the uniquenesses of the two tests. Thus,

$$r_{ab\cdot 123\cdots s} = \frac{\rho_{ab}}{u_a \cdot u_b}$$

or

$$r_{ij\cdot 123\cdots s} = \frac{\rho_{ij}}{u_i\, u_j}, \quad \begin{matrix} i\,, j = a \cdots n \\ i \neq j\,. \end{matrix} \tag{7}$$

In order to avoid the necessity of actually computing the $\frac{n(n-1)}{2}$ residuals as partials, we next seek an expression for the S.D. of the distribution of these partial residuals as a function of the ordinary residuals and the tests' uniquenesses.

Expression (7) defines a variable, the partial residuals for $n(n-1)/2$ intercorrelations, in terms of an index having a variable numerator and a denominator which is the product of two variables. We desire the S.D. of the distribution of this index as a function of statistics determinable from the variables which enter into the index.

It should be noted that since the S.D. of the residuals is usually computed from a double-entry distribution, the mean of the distribution is zero. Similarly the mean of the partial residuals is also zero; that is, if $X_1 = \rho_{ij}$ and $X_2 = u_i u_j$ and we define the index of (7) as $I = X_1/X_2$, then

$$M_I = \frac{1}{n(n-1)} \Sigma \frac{X_1}{X_2} = 0$$

for the double-entry distribution. The distribution variance will therefore take on the form

$$\sigma^2{}_I = \frac{1}{n(n-1)} \Sigma \left(\frac{X_1}{X_2}\right)^2 .$$

The summation includes each partial twice. Obviously, if we work from the distribution of absolute values, with each residual entered once, we have, letting $m = n(n-1)/2$, the following

$$\sigma^2{}_I = \frac{1}{m} \Sigma \left(\frac{X_1}{X_2}\right)^2 = \frac{1}{m} \Sigma \left(\frac{M_1 + x_1}{M_2 + x_2}\right)^2 = \frac{1}{m} \cdot \frac{1}{M^2{}_2} \Sigma\, x^2{}_1 \left(1 + \frac{x_2}{M_2}\right)^{-2} .$$

But

$$\left(1 + \frac{x_2}{M_2}\right)^{-2} = 1 - 2\frac{x_2}{M_2} + 3\frac{x^2{}_2}{M^2{}_2} - \cdots ;$$

whence

$$\sigma^2{}_I = \frac{1}{M^2{}_2}\left[\frac{\Sigma\, x^2{}_1}{m} - \frac{2}{M_2}\frac{\Sigma\, x^2{}_1\, x_2}{m} + \frac{3}{M^2{}_2}\frac{\Sigma\, x^2{}_1\, x^2{}_2}{m} - \cdots\right] . \qquad (8)$$

The 2nd, 3rd, and remaining terms in this expression involve correlations between functions of X_1 and X_2. It seems reasonable to assume that these tend to zero, hence we have

$$\sigma^2{}_I = \frac{\sigma^2{}_1}{M^2{}_2}; \text{ or } \sigma_I = \frac{\sigma_1}{M_2} . \qquad (9)$$

The denominator term of (9) involves the mean of X_2, *i.e.*, the mean of the product of the variables, u_i and u_j. Let $X_3 = u_i$, $X_4 = u_j$, then

$$M_2 = \frac{\Sigma X_2}{m} = \frac{\Sigma X_3 X_4}{m}$$

$$= \frac{\Sigma (M_3 + x_3)(M_4 + x_4)}{m}$$

$$= \frac{1}{m}\left[\Sigma M_3 M_4 + M_3 \Sigma x_4 + M_4 \Sigma x_3 + \Sigma x_3 x_4\right] ; \qquad (10)$$

$$M_2 = M_3 M_4 + r_{34}\sigma_3\sigma_4 \,.$$

But in this case $M_3 = M_4 = M_u$, and r_{34}, the correlation between the m pairs of uniquenesses, should vanish (empirical checks tend to verify this assumption). Hence

$$M_2 = M^2{}_u \,, \tag{11}$$

where M_u is the mean of the uniquenesses for the n tests.

Since the uniqueness for test i is, by definition, $u_i = \sqrt{1 - h^2{}_i}$, we next seek to evaluate M_u in terms of h. Thus

$$M_u = \frac{\Sigma u_i}{n} = \frac{\Sigma\sqrt{1 - h^2{}_i}}{n} = \frac{\Sigma(1 - h^2{}_i)^{\frac{1}{2}}}{n}.$$

Expanding $(1 - h^2{}_i)^{\frac{1}{2}}$, we have

$$(1 - h^2{}_i)^{\frac{1}{2}} = 1 - \tfrac{1}{2}\,h^2{}_i - \tfrac{1}{8}\,h^4{}_i - \tfrac{1}{16}\,h^6{}_i - \cdots .$$

Dropping the 6th and higher order terms,

$$M_u = \frac{\sum\limits_{i=a}^{n} (1 - \tfrac{1}{2}\,h^2{}_i - \tfrac{1}{8}\,h^4{}_i)}{n} \,;$$

$$M_u = 1 - \tfrac{1}{2}\,M_{h^2} - \tfrac{1}{8}\,M_{h^4} \,. \tag{12}$$

But formula (9) calls for M_2, which by (11) is equivalent to $M^2{}_u$. From (12)

$$M^2{}_u = 1 + \frac{1}{4}M^2_{h^2} + \frac{1}{64}M^2_{h^4} - M_{h^2} - \frac{1}{4}M_{h^4} + \frac{1}{8}M_{h^2}M_{h^4} \,;$$

$$M^2{}_u = 1 - M_{h^2} \text{ (approximately)}. \tag{13}$$

Thus, we finally have an approximate value for the S.D. of the partial residuals as

$$\sigma_I = \frac{\sigma_\rho}{M^2{}_u} = \frac{\sigma_\rho}{1 - M_{h^2}}, \tag{14}$$

where σ_ρ is the S.D. of the ordinary residual after s factors have been extracted, and M_{h^2} is the mean communality for s factors.

We are proposing that when σ_I reaches or falls below $1/\sqrt{N}$, the magnitudes of the residuals are such that their departure from zero may be considered as due to chance sampling errors in the original intercorrelations.

Approximation error

The approximation error involved in using (14) instead of the S.D. which would result if all the individual $n(n-1)/2$ residuals had actually been computed as partials by use of (6) should, we believe be small. Consideration of the simplifying assumptions made in arriving at (9) and (10), and the nature of the higher order terms dropped in obtaining (13) should lead the reader to a similar conclusion. In order to have some basis for a better appreciation of the error involved in using (14), we have made a comparison of the S.D.'s yielded by it with the actual values obtained from distributions of partial residuals computed by (6). The results for six checks made on independent sets of data are set forth in Table 1, from which it can be seen that the approximation error is of order .001 in three instances, .002, .003, and .007. This last figure represents an error of about 13% of the actual value, which may or may not be tolerable for a given investigator.

Empirical checks on the criterion

An empirical check on any criterion should be based on setups which not only seem typical of actual experimental situations but which are also designed to permit the operation of certain known facts concerning the sampling behavior of correlation coefficients. Firstly, it must be remembered that the sampling errors in a table of intercorrelations are not independent. This means that one is not justified in adding errors independently to the several r's in a correlation matrix. In the second place, one must not forget that the sampling errors of correlation coefficients tend to yield skewed distributions. This fact has not been taken into account in the published reports. Indeed, one investigator (5) proved by the Chi Squared technique that his injected errors were normally distributed, thereby unwittingly demonstrating that his empirical setup was inadequate. In the third place, an empirical study must, of course, involve the drawing of really random samples from a universe of known factorial description. This, it should be noted, is not the same as adding to individual coefficients chance errors of predetermined variability. The sampling unit in this case must be an individual,* not a slip of paper containing a so-called chance error.

In our empirical series, we have met these requisite conditions by the use of tables of random numbers. This might also be done by

* A person or entity for which measurements are available for the variables being studied by the correlational-factorial method.

the use of coins if it were not for certain practical difficulties involved in tossing the number of coins sufficient for the purpose. A detailed exposition of the manner in which the tables of random numbers were utilized would require too much space, but it should be noted here that the procedure depends upon the defensible assumption that the odd- or even-ness of a digit, as an element, is a chance affair, and therefore analogous to tossing an unbiased coin. Variables may be defined in terms of overlapping or common, plus specific, elements. The theoretically expected r's, obtained by the common element formula, provide the universe correlation matrix of known factorial composition. An "individual," and his scores on the given variables, can be defined in terms of a column of numbers, the column being of predetermined width and yielding scores by an actual count of the odd digits. The limit to the size of a truly random sample is circumscribed only by the extent of available tables of random numbers.

We have tried out the criterion on six independent empirical situations involving true samplings drawn from universes of known factorial composition. All intercorrelations were computed by the product-moment method. The essential facts concerning these setups as regards the size (N) of the samples, the number (n) of variables, and the known number of factors for the universe are given in Table 2, wherein will also be found the limit ($1/\sqrt{N}$) which the given S.D.'s for the partial residuals should reach. It will be noted from this table that the S.D. (σ_I) for the sth partial residuals, where s equals the known number of factors, does in all six cases tend to approximate closely or fall below the criterion level. It will be further noted that the reduction in the S.D. for the partial residuals which results from extracting $s + 1$ factors is very small, especially when compared to the reduction in S.D. as one goes from $s - 1$ to s factors.

We have tried out the proposed criterion on Brown and Stephenson's (2) data, for a sample of 300 and the 19 variables left after the purge of tests which disturbed the single factor hypothesis. These data, it will be recalled, satisfied the tetrad criterion. The S.D. of the first partial residuals is .064, which is not quite down to the value of $1/\sqrt{N}$, *i.e.*, .058. Application of the criterion to the data of Thurstone (6) based on 57 tests and an N of 240 would indicate the extraction of fewer than five factors. The fact that Thurstone has reported additional data on new samples which tend to confirm the existence of 7 or 8 factors for his battery would seem to cast considerable doubt on the validity of our proposed criterion. It must be noted, however, that the consideration of the effect of chance sampling errors in Thurstone's data is subject to a handicap—the actual number of individuals taking the various tests range from 104 to 234, and we are nowhere told just what the *N*'s are for the intercorrelations. Presumably they will be still lower, and most certainly variable.

Check on other criteria

It may not be out of order to indicate how well a couple of other proposed criteria for the number of factors work on our six empirical series. The first criterion of Tucker (**6**, p. 66) and later versions thereof (**1**, **7**) fail in all six cases. The more recently proposed criterion of Coombs (**3**) checks in the 9-1 and 12-2 series, but for the remaining four series, one would need to extract at least two or more factors than exist in the universe before the criterion is reached. Strangely, for the 15-3 setup, a strict adherence to the Coombs criterion would not permit the extraction of *any* factors even though it is known that three sizeable factors exist in the universe from which the sample was drawn.

The failure of the Tucker criterion need not be surprising—the surprising thing would be to find that it did work since it is solely a function of n, the number of variables. The present writer cannot entertain any hopes for Coombs' proposal since it also is mainly a function of n. Surely the size of the sample must have something to do with the amount of variance which may be attributed to chance sampling.

Although our proposed criterion seems to work fairly satisfactorily, there are two limitations which should be mentioned. Firstly, we have no evidence that it will be adequate for those situations, most frequently encountered in practice, where the number of variables is larger than in our empirical series. And secondly, we must admit that the sampling error of the residual, $r_{ab\cdot 12\cdots s}$, as a partial correlation may not be strictly comparable to that of the ordinary partial since r_{ab} has been utilized in determining the values of the factor loadings which have then been used in calculating the partial residuals. These points need further investigation.

TABLE 1

Empirical Check on Error Involved in Approximation By Formula (14) for the S.D. of the sth Partial Residuals (Six independent sets of data)

No. variables	9	10	10	10	12	15
s	1	2	3	4	2	3
Actual S.D.	.086	.073	.052	.047	.060	.069
By formula (14)	.088	.074	.045	.046	.061	.066

TABLE 2

Empirical Check on the Partial Residual Criterion

N	150	208	250	208	150	150
No. variables	9	10	12	10	15	10
No. factors	1	2	2	3	3	4
$1/\sqrt{N}$	.082	.069	.063	.069	.082	.082
${}_1\alpha_I$	.088*	.468	.364	.301	.343	.221
${}_2\sigma_I$	.081	.074*	.061*	.198	.288	.154
${}_3\sigma_I$		.066	.055	.045*	.066*	.095
${}_4\sigma_I$				.041	.061	.046*
${}_5\sigma_I$						.043

* These values should be down to or below the values given for $1/\sqrt{N}$.

REFERENCES

1. Blakey, R. A re-analysis of a test of the theory of two factors. *Psychometrika,* 1940, 5, 121-136.
2. Brown, W., and Stephenson, W. A test of the theory of two factors. *Brit. J. Psychol.,* 1933, 23, 352-370.
3. Coombs, C. H. A criterion for significant common factor variance. (Unpublished)
4. Kelley, T. L. Statistical method. New York: Macmillan, 1923.
5. Mosier, C. I. Influence of chance error on simple structure, etc. *Psychometrika,* 1939, 4, 33-44.
6. Thurstone. L. L. Primary mental abilities. *Psychometric Monograph* No. 1, 1938.
7. Wright, R. E. A factor analysis of the original Stanford-Binet scale. *Psychometrika,* 1939, 4, 209-220.

PSYCHOMETRIKA—VOL. 7, NO. 1
FEBRUARY, 1942

ITEM SELECTION BY THE CONSTANT PROCESS*

GEORGE A. FERGUSON
DEPARTMENT OF EDUCATIONAL RESEARCH, UNIVERSITY OF TORONTO

This paper relates the constant process used in psychophysics to the problem of item selection. Each test item may be described in terms of a limen, which is an index of the point at which an item discriminates, and the standard deviation of the limen, which is an index of the 'goodness' of discrimination. The method developed may be related not only to the description of items but also to the description of persons. Thus a person's ability may be described in terms of a limen and its standard deviation.

Within recent years an increasing tendency has become apparent among psychometricians to bring the methods of mental measurement more directly in line with the psychophysical methods of experimental psychology. This tendency is particularly marked in existing methods for the scaling and standardization of tests and in some of the many techniques recently developed for the selection of test items. The application of the psychophysical methods in the scaling and selection of test items was foreshadowed by a number of writers: Binet (1, 1908), Thurstone (2, 1925), Thomson (3, 1926), and Symonds (4, 1929). Guilford (5, 1936) furnished a complete formulation of the problem with which this paper concerns itself, but offered no solution. He writes, "If one could establish a scale of difficulty in psychological units, it would be possible to identify any test item whatsoever by giving its median value and its 'precision' value in terms of h as in the method of constant stimuli. This is an ideal towards which testers have been working in recent years and already the various tools for approaching that goal are being refined."

The only practicable solution to this problem as formulated by Guilford involves the establishment of an arbitrary scale of ability on the assumption that ability is normally distributed in the population, and the description of the performance of any given person in terms of σ-units on this arbitrary scale. This is an orthodox statistical procedure, and is frequently used in the standardization of tests. The procedure is valid when the age-range of persons tested is small. When, however, the age-range is large a development of the technique

* I am indebted to Godfrey H. Thomson of the University of Edinburgh for reading and criticising part of this paper, and to D. N. Lawley of the same University for assistance in the development of certain of the arguments contained herein.

is required since the variance of ability is not independent of age. In the following discussion we shall confine ourselves to a single year of age-range, making the assumption that the variance of ability is equal for each month of age.

In the estimation of a two-point tactual limen by means of an aesthesiometer, for example, the limen is regarded as that point, usually in millimetre units, where the probability of either a one-point or two-point judgment is one half. The scatter of the limen is described in terms of its variance or in terms of a "precision value," h, which is in fact a *weight* proportional to the reciprocal of the standard deviation of the limen. In the application of the constant process* to item selection, we define the limen as that point measured in "σ-units" of ability where the probability of a person of that ability either passing or failing the item is one half. This limen is the point of discrimination. The standard deviation of the limen is an indication of the "goodness" of discrimination. Having estimated these two parameters for each item of a test we are in a position to estimate the probability of any given person passing any given item. The estimation of such a probability is fundamental in the reduction of mental-test method to a sound theoretical basis. Within recent years much effort has been wasted in the effort to apply correlation methods to problems where the more elementary theory of mathematical probability would have produced results of substantially greater simplicity and value.

In outlining the application of the constant process to item selection let us presume that we have constructed a provisional test of a large number of items from which we wish to select those items which are to be included in the finished test, and that we have given this provisional test a preliminary tryout on a representative sample of the age range for which the test is ultimately intended. Let us presume that the age range is not greater than a single year. The score obtained by a person on the complete test must, in the absence of any better criterion, be regarded as the best available estimate of a function of that person's ability.

On the assumption that the distribution of ability is normal, let us divide the persons tested into k categories, the class interval expressed in "σ-units" of each category being the same. Thus if we were to divide our persons into seven categories ($k = 7$) and to adopt .60 as the class-interval, the percentage falling in each category would be as shown in Figure 1.

Assign to each category a value x in terms of "σ-units" equal to

*Throughout I have adhered to the practice suggested by Thomson (6) of using the word *process* to imply a process of calculation.

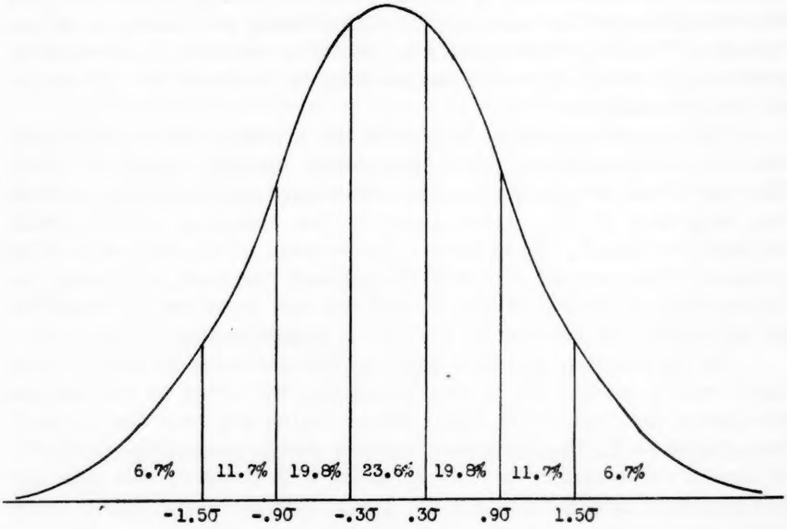


Fig. 1

the mid-point of that category. For all practicable purposes it will be sufficient to regard the mid-points of the categories at the tails of the distribution as in the above example -1.8σ and 1.8σ. We have now grouped our sample of persons into categories on a scale of ability in which the differences between the mid-points of the several categories are considered equal.

We now construct an answer-pattern, and determine the number and proportion of persons in each ability category passing each item. The proportions, $p_1, p_2, \cdots, p_k$, when plotted against the mid-points, $x_1, x_2, \cdots, x_k$, of the corresponding ability categories expressed in "σ-units" should, on the basis of the phi-gamma hypothesis, form the integral of the normal curve of error; that is, if we transform the proportion p to a new variable y by the relation

$$p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{u^2}{2}} du,$$

then on the basis of the phi-gamma hypothesis there exists a linear relationship between x and y.

From the proportions $p_1, p_2, \cdots, p_k$, the value of the limen and its standard deviation can be estimated for each item in terms of "σ-units" of ability. These two parameters describe the functioning of the item, the limen being a measure of the point at which the item discriminates and the standard deviation being an indication of the "goodness" of discrimination. Any one of a number of processes in general use among psychophysicists may be used for the estimation of these parameters.

For a complete rationalization of the problem, the *constant process,* sometimes termed the Müller-Urban method, should be used. The use of this process as applied to this particular problem involves the weighting of the observations by the combined Müller-Urban weights and also by the number of cases upon which each value of p is based. This process, although theoretically the most admirable, involves much arithmetical labour, and can not, therefore, be regarded as practicable for the routine purpose of item selection.

The application of Spearman's arithmetic-mean process is to a large degree invalidated in this connection by virtue of the tail assumptions involved. With items whose limina are near the mean of the ability scale, the Spearman process yields reasonable estimates of limina and standard deviations, since with items of this type the tail assumptions necessary are not great. However, with items whose limina are near the extreme of the ability scale, it becomes necessary to estimate the required parameters from a few values of p at one end of the distribution. With items of this type it becomes impossible to fix in any reasonable fashion the tails of the distribution.

Other processes suggest themselves which are simple in type and yield results sufficiently satisfactory for routine purposes. The first is the process of simple linear interpolation in which the median or 50% point is regarded as an estimate of the limen. This process has the disadvantages that it uses only two values of p, that it involves the assumption that the curve is a straight line between the two values of p in question, and that it yields no measure of scatter. The last of these objections may be eliminated by calculating the 16% and 84% points, and taking half the distance between these two points as a rough estimate of the standard deviation. With items that are very difficult the 84% point, and with items that are very easy the 16% point, can not readily be calculated. In such cases it is necessary to use the difference between either the 84% point or the 16% point and the 50% point as a very rough estimate of the standard deviation. While this process is open to many serious objections on theoretical grounds, the calculation involved is small, and the estimate obtained, although subject to substantial error, will be found satisfactory when no great accuracy is desired.

A process that avoids some of the disadvantages of linear interpolation between the observed proportions is to determine a sigma value y corresponding to each value of p in the normal ogive from tables prepared for this purpose. The sigma values y when plotted against values of x should assume a linear relationship. We can now proceed to estimate the required parameters either by simple interpolation between the sigma values or by calculating the slope of the best-fitting least-squares line. This least-squares line process is in fact the constant process without the use of weights. If the data are arranged such that $\sum x = 0$, the slope of this line is given by the relation

$$b = \frac{\sum x y}{\sum x^2},$$

and the limen by the relation

$$L = \frac{\sum y}{k b}.$$

The standard deviation of the limen is equal to the reciprocal of the slope.

To illustrate the functioning of item selection by the constant process and to determine the relative efficacy of the various processes suggested for estimating the limen and standard deviation, the following short experiment was conducted. From the test scripts of a complete year group of 11+ children, who had taken a Moray House Test, a sample of 216 scripts chosen at random was selected. The 216 children whose scripts were selected were found to be representative of the complete year group of 11+ children. These 216 scripts were then divided into seven categories of equal class interval in terms of "σ-units." The following table shows the number and proportion of persons in each category.

The mid-points of the categories at the tails of the distribution

TABLE 1

Mid-point of class interval x (σ-units)	% in category	No. in each category
1.8	6.7	15
1.2	11.7	25
.6	19.8	43
0	23.6	50
− .6	19.8	43
−1.2	11.7	25
−1.8	6.7	15

TABLE 2

x (σ-units)	-1.8	-1.2	$-.6$	0	.6	1.2	1.8
item	Proportions passing item						
1	.33	.56	.70	.94	.95	.96	.93
2	.13	.36	.65	.82	.91	1.00	1.00
6	.07	.00	.30	.44	.77	.88	1.00
12	.07	.20	.28	.52	.63	.88	.87
20	.13	.12	.12	.28	.47	.72	.87
28	.00	.08	.05	.18	.30	.36	.47

are taken as 1.8σ and -1.8σ. These values are of course not the midpoints, but the error introduced by their adoption is negligible.

An answer pattern was then constructed and the number and proportion of persons in each category passing each item were determined.

For purposes of illustration, Table 2 gives the proportion of persons in each category for six different items.

From these proportions we may proceed to estimate a limen and a standard deviation for each item. For comparative purposes these parameters have been estimated for the foregoing six items in a number of ways.

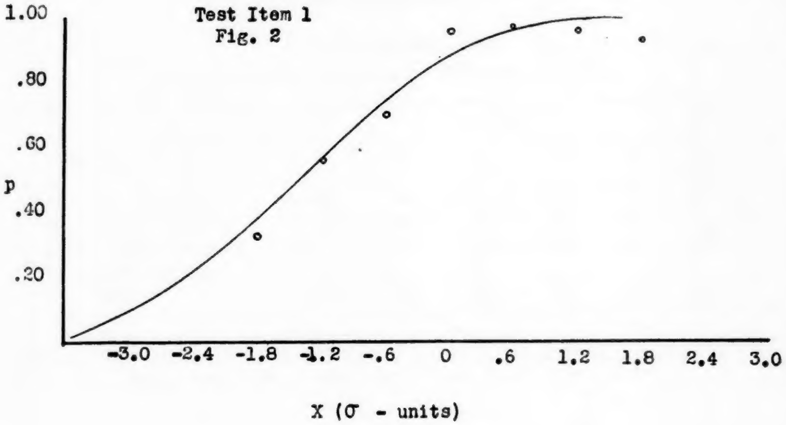
Firstly, the constant process,* or Müller-Urban method, was used. Each value of p was weighted by the Müller-Urban weights and by the number of observations upon which it was based. This amounts to weighting each proportion p by the product of the Müller weight and the quantity $\frac{N}{p\,q}$, the reciprocal of the variance of each proportion. The values of the limina calculated by this process are given in Table 3 and the corresponding standard deviations in Table 4. Figures 2 to 7 show the obtained values of p plotted against values of x for the six items with the best fitting normal ogives.

The limina and standard deviations were calculated also by Spearman's arithmetic-mean process, by simple linear interpolation, and by fitting a least-squares line to the sigma values y corresponding to values of p in the normal ogive (the constant process without weights). The values of the limina calculated by these three pro-

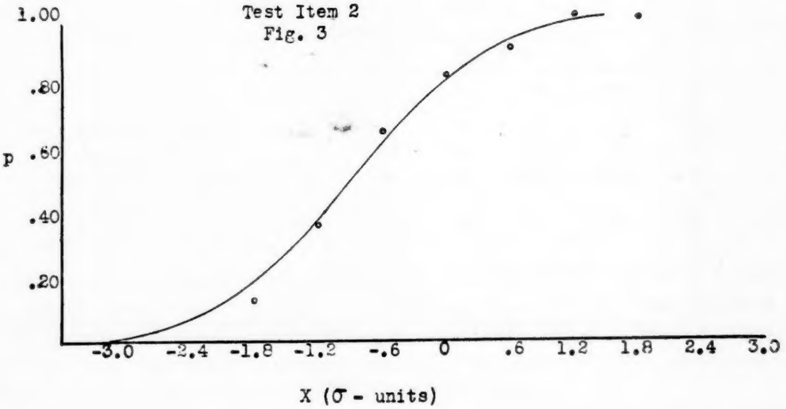
* Although the term constant process is correct, in the strictest conventional sense, only when limina and precision values are estimated, I employ the term to relate also to the process whereby standard deviations are estimated. The standard deviation is a simple function of the precision value, the relation being

$$s^2 = \frac{1}{2h^2}.$$

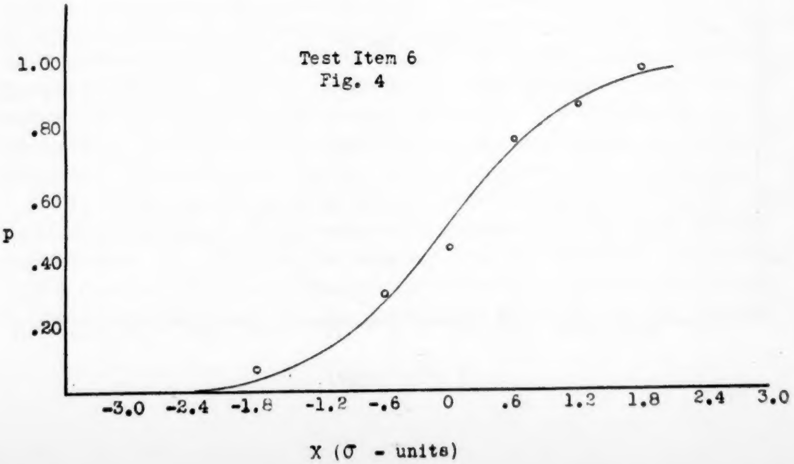
Throughout I have calculated values of s instead of values of h.



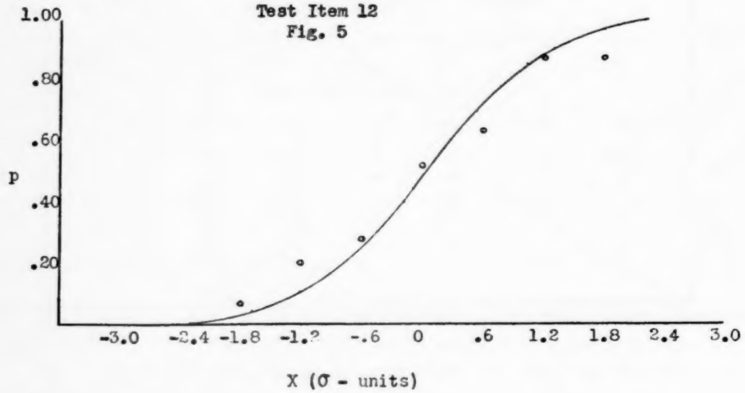
Test Item 1
Fig. 2

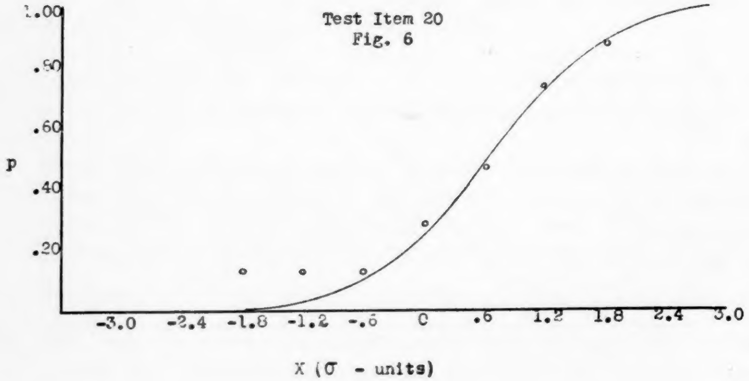


Test Item 2
Fig. 3



Test Item 6
Fig. 4


Test Item 12
Fig. 5


Test Item 20
Fig. 6

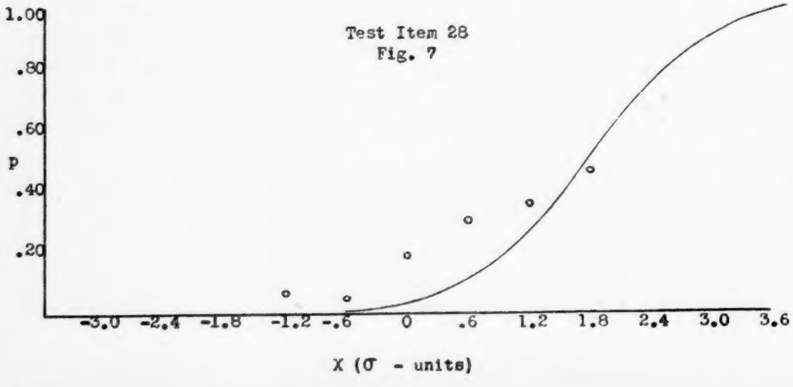

Test Item 28
Fig. 7

TABLE 3

Values of Limen Obtained by Different Processes

	Test item					
Process	1	2	6	12	20	28
Constant Process	—1.48	— .88	— .03	.06	.64	1.79
Arithmetic Mean	......	— .82	.03	.07	.47	......
Linear Interpolation	—1.36	— .91	.11	— .05	.67	1.96
Least squares line process	—1.58	— .82	.03	— .04	.51	2.12

cesses are given in Table 3, and the values of the corresponding standard deviations in Table 4.

TABLE 4

Values of Standard Deviations Obtained by Different Processes

	Test item					
Process	1	2	6	12	20	28
Constant Process	1.44	1.00	1.05	1.31	1.40	1.88
Arithmetic Mean	......	.86	.89	1.22	.87	......
Linear Interpolation	1.11	.93	1.16	1.24	1.29	2.06
Least-squares line process	1.66	.93	.93	1.31	1.48	1.87

Spearman's arithmetic-mean process is invalidated for items of types 1 and 28 by virtue of the tail assumptions involved. If the constant process can be taken as furnishing the best estimates of the required parameters, Spearman's process, as is to be expected, furnishes substantial underestimates of the standard deviations.

The estimates obtained by fitting a least-squares line to the sigma value y corresponding to values of p in the normal ogive differ somewhat from the estimate obtained by the constant process. These differences result from the absence of weights. Values obtained without the use of weights must be regarded as rough estimates only. In applying the least-squares line process, it will be found desirable to delete certain of the extreme values of y which are based on comparatively few cases and to fit the line to the four or five central points.

The estimates obtained by linear interpolation seem to approximate about as closely to the estimates obtained by the constant process as those obtained by the least-squares line process. When the data neither warrant nor demand great accuracy, the results obtained by linear interpolation will be sufficiently approximate for routine purposes. If substantial accuracy is desired, a sample much greater than 216 cases is necessary.

The application of the method described above relates mental-testing technique directly to the psychophysical methods. From estimates of the limen and standard deviation it is possible to estimate the probability of any given person passing any given item. For example, the limen and standard deviation for item 20 estimated by the constant process are respectively .64σ and 1.40. This limen implies that the probability of a person of ability .64σ either passing or failing the item is $\frac{1}{2}$. The probability of a person of ability 2.04σ passing this item is .84, while the probability of a person of ability $-.76\sigma$ passing the item is .16. I am of the opinion that the estimation of such probability will have further application in mental-test theory.

Just as we describe an item in terms of a limen and standard deviation, it is possible to reverse the process and, on the assumption that the items on a given test are representative of a defined population of items, to describe each person in terms of a limen and a standard deviation. Thus the ability of a person could be described in terms of a level of difficulty where the probability was $\frac{1}{2}$ that a person would either pass or fail tests of that difficulty.

Such a limen would be closely analogous to a stimulus threshold. Furthermore, we could describe the relative abilities of a number of persons in terms of the probabilities of their passing a task at an arbitrarily specified level of difficulty.

The technique of constructing a test whereby persons could be described in terms of limina and standard deviations or precision values would seem at the moment to be something as follows. As previously we divided persons tested into k categories with equal class intervals expressed in "σ-units," so now a test could be constructed of k subtests of increasing order of difficulty. Our subtests would be selected such that the differences in difficulty between them would be in terms of equal "σ-units" of difficulty, relative, of course, to some defined population. One would then proceed to calculate for any given person the proportion of his successful responses in each subtest, ascending as they do in equal "σ-units" of difficulty, and assume as before that the proportions, $p_1, p_2, \cdots, p_k$, when plotted against the mid-points, $x_1, x_2, \cdots, x_k$, of corresponding *difficulty categories,* form the integral of the normal curve of error.

The next step would be to estimate for each person a limen and a standard deviation or precision value by any of the methods described above or for that matter by any efficient method of estimation. The limen for persons would be in terms of "σ-units" of difficulty. Thus, ability is described in terms of a parameter the implications of which is that the probabilities are equal that any given person will either pass or fail tasks of a certain defined difficulty.

In such a procedure we may well enquire what would be the meaning of our second parameter for persons, the standard deviation or precision value. It is apparent that this second parameter is an index of the degree to which the responses of any given person differ from the response of the average person in a defined population. Thus, what person *a* finds difficult may differ somewhat from what person *b* finds difficult and from what is regarded as difficult by the average person in a given population. Thus, although the average ability, as it were, of a given person over a series of tasks may be about the same as that of the average person, he may find some tasks easier and some more difficult than the average person; that is, difficulty for any given person is not the same as difficulty for the average person. Hence our second parameter is a measure of a difference between any given person and the average person, and as such is a *type parameter*. Indeed it bears a kinship to the second or bipolar factor usually found in the factorial analysis of persons.

REFERENCES

1. Binet, A. and Simon Th. *Ann. Psychol.*, 1908, 14.
2. Thurstone, L. L. A method of scaling psychological and educational tests. *J. educ. Psychol.*, 1925, 16, 433-451.
3. Thomson, G. H. A note on scaling tests. *J. educ. Psychol.*, 1926, 17, 551-553.
4. Symonds, P. M. Choice of items for a test on the basis of difficulty. *J. educ. Psychol.*, 1929, 20, 481-493.
5. Guilford, J. P. Psychometric methods. New York and London: McGraw-Hill Book Company, 1936.
6. Thomson, G. H. A direct deduction of the constant process used in the method of right and wrong cases. *Psychol. Review*, 1919, 26, 454-464.

PSYCHOMETRIKA—VOL. 7, NO. 1
FEBRUARY, 1942

THE MEASUREMENT OF CONFORMITY

E. T. KATZOFF
NORTHWESTERN UNIVRESITY

Allport's *J*-curve hypothesis of conforming behavior and its attendant treatment of appropriate data are criticized on the following points: (1) narrowness of application; (2) flexibility of interpretation of results; (3) arbitrary selection of a criterion of conformity; (4) lack of a means by which the extent of conformity in one situation can be compared with that in another; (5) inequality of "telic" units. As an alternative treatment of such data, the method of higher moments is suggested and rationalized. Data from Allport and Solomon are reworked by this method and results compared.

F. H. Allport (2) has indicated the importance of conformity to sociology and social psychology. We agree that there is a need for more quantitative studies in this field and for a methodology. The validity of the "*J*-curve hypothesis" in this respect is, however, open to question. It is the purpose of this paper to investigate some of the shortcomings of this hypothesis and to suggest an alternative method based on established statistical procedures.

The "*J*-Curve Hypothesis"

Fields of Conformity. Basic to the "*J*-Curve hypothesis" is what Allport calls a "conformity situation" or "field of conformity" which he has defined (2, p. 912) as follows: "A conformity field exists when there is a generally accepted, though not necessarily explicitly stated, rule and purpose in the situation, and when *fifty per cent** or more of the population fall upon the first step of a telic (conformity) continuum whose variable is degrees of fulfillment of this purpose."

Without such a "field of conformity," the "*J*-curve hypothesis" is not applicable. Thus this approach is limited to cases where (1) there is a discoverable, though "not necessarily explicitly stated" rule or custom; (2) "complete conformity" has been defined and (3) a majority of the population "conform completely."

If we assume that the "*J*-curve hypothesis" can be used to measure conformity where it exists, it follows that no conformity of behavior obtains unless the foregoing conditions hold. If on the other hand we hold that there may be conformity of behavior even in the event that we are not immediately able to define a specific rule or cus-

* Italics mine.

tom or where somewhat less than 50% of the population conform fully, it immediately follows that even if valid, the "*J*-curve hypothesis" has at best a very limited application. A more widely applicable technique for the study of conformity is needed.

Having defined the "field of conformity," Allport continues (2, p. 913):

> If, in any field of conformity (see definition given above) we apply a scale whose steps are variations of behavior which represent successive recognizable degrees of fulfillment of the "accepted common purpose," ranging from the prescribed or "proper" act, which most completely fulfills the purpose (on the left) to that which gives it the least recognizable amount of fulfillment (upon the right), the following will occur: (a) more instances will fall upon the step at the extreme left than upon any other; (b) the successive steps from left to right will have a respectively diminishing number of instances; and (c) the decline in the number of instances will decrease as we proceed by successive steps from left to right.

A second portion of the hypothesis dealing with the empirical continuum is stated as follows (2, p. 916-917):

> In any conformity field the distribution of measurable variations of the behavior upon a relevant empirical, or nontelic, continuum is in the form of a . . . . unimodal, double-*J*-curve (i.e., a curve having positive acceleration of both slopes), in which the mode is likely to be off center and the slopes are likely to be asymmetrical.

We shall consider each of these parts of the statement of the hypothesis separately and in reverse order.

The "Double-J" Curve. In the earlier descriptions of the form of the empirical distribution of conformity data, Allport (1) stressed three points. First, the curve was said to be steep or leptokurtic, second, positively accelerated toward its single mode from both directions, and third, "likely" to be asymmetrical.

Dudycha(5) pointed out that in applying statistical measures of kurtosis to his own as well as to Allport's data, some of the distributions were found to be leptokurtic, but others more mesokurtic and some even platokurtic. In reply to this, Dickens and Solomon (4) claim that Allport had misused the term "leptokurtic" and that "the distinction between a normal distribution and a "double *J*" is based chiefly on the criterion of acceleration as you approach the mode." From this it appears that the "double-*J*" may be steep or flat, symmetrical or skewed, so long as it is positively accelerated toward the mode from either direction.

The normal curve is positively accelerated toward the mean (or mode) from both directions up to one standard deviation either side of the mean. If it so happened that in studying any particular distribution an investigator happened to select a rather gross step interval, a perfectly normal distribution would take on the most distinguishing characteristic of the "double-*J*." To verify this possibility we have taken a distribution of the height of 1079 men as presented by Pearson and Lee (7). A chi squared test of goodness of fit indicated that any deviation of this distribution from normal must be attributed to chance factors. When these data are distributed with the use of a three-inch step interval, the slopes of the resulting curve are posi-

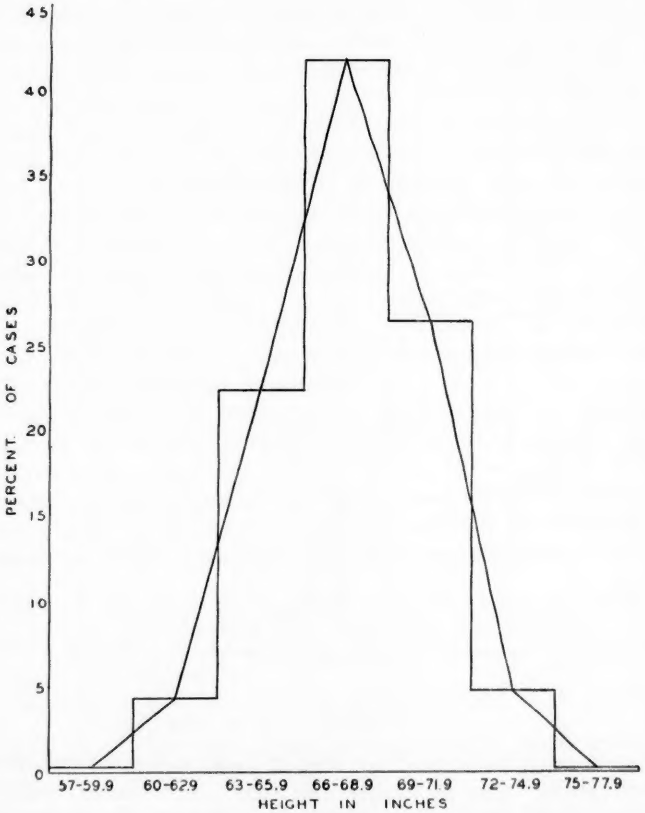


FIGURE 1

tively accelerated toward the mode throughout the range (Fig. 1). Thus can perfectly normal data yield a perfect "double-*J*" curve.

The "telic continuum." In constructing a "telic continuum" it is first necessary to define the purpose being fulfilled The next step would be to determine what is to be considered "full conformity." One question that immediately arises in this respect concerns the extent to which behavior reflects the intent of the individual. It is impossible to say because one is late to work that he did not intend to be on time.

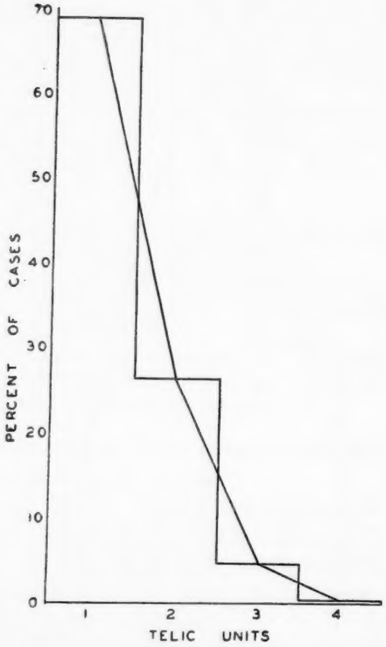


FIGURE 2

Further, it must be pointed out that in many cases the definition of "full conformity" may be highly arbitrary. A "*J*-curve" will result from almost any distribution if the proper left-hand step is chosen. Thus if we refer to the data on the height of 1079 men (see above) and say that people should not be more than 69 inches tall, we may construct a "telic continuum." This continuum is presented in Figure 2. The original distribution of the data on which it is based, it will be recalled, is normal.

"Telic" Units.

If the "telic continuum" is to be of any use at all, units on it should be at least psychologically if not physically equal. In early studies such as that of conformity of the response of drivers to a stop sign at an intersection, little attempt was made to equate the intervals. In this study, [Allport (1)] the units employed were (a) come to a complete stop, (b) slow down considerably (c) slow down slightly, and (d) go ahead with no reduction of speed. If we consider a driver who had been driving at 50 miles an hour and slowed down to 25, another who had been going 30 and slowed to 25, and a third who had been going 25 and did not reduce his speed at all we have three individuals all crossing the intersection at the same speed. From one point of view, each of these drivers is failing to conform to the same extent. Yet according to Allport's criteria, each would fall on a different step of the continuum. By following Allport's logic in this case, if all three drivers had come to a complete stop, we would have to consider the first as conforming more completely than the other two, since his decrease of speed would have been greatest. By the same token, the second must in stopping have conformed to a greater extent than the third. Thus complete conformity would require not only that the driver stop, but that the first obtain the greatest possible speed. This *reductio ad absurdum* can be applied equally well to other measures of "complete conformity."

In the study of Allport and Solomon (3) on lengths of conversation in church, library, and club-room situations, an attempt was made to obtain equal "telic" units. To obtain an evaluation of various degrees of annoyance due to conversation of other people in such situations, five statements were selected for each of the three situations. These statements for the library situation were:

1. I am slightly distracted from what I am doing.
2. I am beginning to notice the conversation.
3. I feel like stopping my activities and staring or actually do so.
4. I feel like getting up and asking them to stop or actually do so.
5. I feel like asking an authority to make them stop or actually do so.

The extent of annoyance expressed by each of these feelings or actions was rated by a modification of the Thurstone attitude technique, and an annoyance value was assigned to each.

Next, a scale was prepared for each of the situations. On each scale each of the five statements appeared. Below each was a line marked off in 15-second intervals. The subjects were asked to indicate by checking on the appropriate line how much time would elapse before each thought he would experience the state of annoyance expressed by the statement.

A graph was then constructed for each of the situations. Along the abscissa were plotted the median times at which the individuals tested reported that they would experience the state described by each of the five statements rated. Along the ordinate, the annoyance values of each of the five statements were plotted. Horizontal lines were drawn from each of the five points on the ordinate to where each met a vertical line through the corresponding (i.e., from the "time value" of the same statement) point on the abscissa. By drawing a line through the points at which the five sets of lines intersected, a continuous and infinitely divisible unit of time with respect to equal units of annoyance value was graphically determined.

This aspect of the experiment was duplicated with respect to the library situation in the psychology classes at Northwestern University.* In addition to the 15-second interval used by Allport and Solomon, intervals of 60 seconds, 5 seconds, and one second were employed. If the results obtained by Allport and Solomon were independent of the interval employed, any change in the interval would not influence the size of the resulting units. Table 1 gives a comparison of the present results with those obtained by Allport and Solomon. Where the intervals were both 15 seconds, the results were comparable. Where different time intervals were used, the results were entirely different. Table 2 shows the "infinitely divisible telic units" as based on three different time intervals. Obviously the units arrived at by Allport and Solomon reflected primarily the time interval em-

TABLE 1

Comparison of the Appearance of Different Degrees of Annoyance as a Function of the Time Interval Employed. All Data in Library Situation.
(All Measurements are in Minutes)

Degree of Annoyance	Allport & Solomon		Northwestern Data		
	1/4 min.	1/4 min.	1 min.	1/12 min.	1/60 min.
A	1.25	2.00	4.63	1.00	.25
B	1.58	2.00	3.67	.51	.20
C	3.15	2.50	4.34	1.00	.33
D	5.75	4.66	14.48	3.00	.50
E	7.90	7.53	25.00	5.00	3.00
N	56	40	118	69	104

* These data were originally presented before the 1940 meetings of the Midwestern Psychological Association.

TABLE 2

Comparison of Equally Divisible Telic Units as a Function of the Time Interval Employed
(All Measurements are in Minutes)

Telic Step	Allport & Solomon	Northwestern Data	
	1/4 minute	1/12 minute	one minute
Complete Conform.	1.5	.20	3.75
1	1.9	.225	4.00
2	2.4	.250	4.25
3	2.8	.275	4.50
4	3.3	.300	6.00
5	3.8	.350	7.87
6	4.3	.375	9.87
7	4.85	.400	11.87
8	5.40	.425	17.50
9	6.15	1.32	22.00
10	6.15	2.65	

ployed by them. Since the results are a function of the scale employed rather than the general situation, they are of questionable significance.

Having considered certain inadequacies of the *J*-curve hypothesis in its present form, it is a further purpose of this paper to re-investigate the phenomena of conformity both from theoretical and empirical points of view in order to determine a more adequate method of treating the data.

Quantitative Comparisons of Conformity. One of the greatest weaknesses of the "*J*-curve hypothesis" lies in the fact that it provides no means by which the extent of conformity can be given quantitative expression. A chi squared test may be made to determine whether the empirical observations differ significantly from a normal distribution. No similar technique, however, exists for the testing of the significance of a single "*J*-curve" or for determining whether greater conformity exists on one "*J*-curve" than on another.

Both Dudycha (5) and Solomon (8) have attempted to give a mathematically expressed index of conformity. Since, however, neither has produced the distribution function for his formula and its moments, the expressions are rendered useless by virtue of the fact that they provide no means by which to determine whether an obtained difference between two "*J*-curves" could have arisen by chance.

Any measure of conformity, if it is to be at all useful in the study of social psychology, must be able to indicate whether or not in observed data conformity to the extent observed could be attributed to chance, and whether differences between conformity in two independent situations could have so arisen. It is our hope to describe a method of measuring conformity which is free from the artifacts of the "*J*-curve hypothesis" and at the same time exercises the abovementioned controls.

A Statistical Theory for Measuring Conformity

Basic Tenets. It is generally acknowledged that the final appearance of a bit of behavior is the result of the interplay of a great number of factors. Some of these tend to promote the appearance of the behavior while others tend toward its inhibition. Some of the factors are strong and some are weak. We may never know all of the factors that finally result in a given bit of behavior, but without some such theory of causality the entire world of science as we know it today would collapse.

If we were to label each of the many factors which work to bring about the appearance of any bit of behavior under consideration as p_j where $j = 1, 2, 3, \cdots, k$; and if we were to label all the inhibitory factors as q_u where $u = 1, 2, 3, \cdots, v$, then by setting the total sum of these factors equal to unity, we would have the relative weight of each; thus,

$$\sum_{u=1}^{u=v} q_u + \sum_{j=1}^{j=k} p_j = 1. \tag{1}$$

The first term on the left-hand side of equation (1) gives us the relative weight of all the inhibitory factors while the second term gives us the relative weight of the facilitory factors. If we call these q and p respectively, then (1) reduces to

$$q + p = 1. \tag{2}$$

Thus q comes to represent the probability that the given behavior will fail to appear in any given instance and p the probability of its appearance. By raising this expression to the sth power, thus,

$$(q + p)^s = 1^s, \tag{3}$$

we achieve the probability distribution of the occurrence of the behavior.

When s is large and the factors favoring the appearance of the behavior are equal to those opposed to it, the binomial expansion above approaches a normal curve; thus,

$$\lim_{s \to \infty} (q + p)^s \equiv (2\pi)^{-\frac{1}{2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx, \tag{4}$$

where

$$q = p = \tfrac{1}{2}.$$

The moments of the binomial in terms of those of the normal curve are

$$\bar{x} = sp; \tag{5}$$

$$\sigma = \sqrt{s\,p\,q}; \tag{6}$$

$$\alpha_3 = \frac{q - p}{\sqrt{s\,p\,q}}; \tag{7}$$

$$\alpha_4 = \frac{1}{s\,p\,q} - 6/s + 3. \tag{8}$$

Every psychologist is familiar with the first two moments, the mean ($\bar{x}$) and the standard deviation (σ). α_3 and α_4 are the third and fourth moments about the mean of a distribution of standard measures. They may be computed from obtained data as follows:

$$\alpha_3 = \frac{1}{N} \sum_{x=1}^{x=n} \left(\frac{x - \bar{x}}{\sigma} \right)^3 \tag{9}$$

and

$$\alpha_4 = \frac{1}{N} \sum_{x=1}^{x=n} \left(\frac{x - \bar{x}}{\sigma} \right)^4. \tag{10}$$

For the normal curve the values of these parameters are

$$\bar{x} = 0; \tag{11}$$

$$\sigma = 1; \tag{12}$$

$$\alpha_3 = 0; \tag{13}$$

$$\alpha_4 = 3. \tag{14}$$

Application of α_3 to conformity.

The assumption underlying the application of the foregoing binomial to data is that the factors making for the occurrence of the behavior under consideration are equal in potency to those which would make for the failure of the behavior to occur. In a situation where social or other forces operate strongly to bring about a certain type of behavior, the equality of p and q no longer exists and $p > q$. Where the forces operate to inhibit a mode of behavior, the opposite relationship obtains and $p < q$.

In cases where $p = q = \frac{1}{2}$ the curve is symmetrical and $\alpha_3 = 0$. When $p \neq q$ the symmetry no longer exists and $\alpha_3 \neq 0$.

From the formula

$$\alpha_3 = \frac{q - p}{\sqrt{s\,p\,q}},$$

it is clear that the size of α_3 will vary with the difference between p and q when s is constant. Thus to test whether or not factors are at work to bring about conformity in a given social situation would require testing the null hypothesis that the a_3 of the distribution does not differ from zero. If this hypothesis is rejected[*] it may be stated, without recourse to any of the artifacts of the "*J*-curve hypothesis," that factors are operating which tend to force conformity. Further, it is possible by the use of this statistical treatment to determine whether the conformity in one situation is significantly greater than that in another. This may be accomplished by taking the ratio of the difference of the α_3's to the standard error of that difference. If we call this ratio "*t*," the formula for its determination is

$$t = (\alpha_{3_a} - \alpha_{3_b})\,(\sigma^2{}_{a_{3_b}} + \sigma^2{}_{a_{3_a}})^{-\frac{1}{2}}; \tag{15}$$

where

$$\sigma^2{}_{a_3} = \frac{6N(N-1)}{(N-2)\,(N+1)\,(N+3)}. \tag{16}$$

Conformity vs. Uniformity.

Allport and Solomon (p. 420) point out the fault of failing to distinguish between conformity and simple uniformity. They do not, however, give any statistical measure with which to determine this difference. They do assume that uniformity is greatest in the church situation by virtue of the fact that 27% of the cases fall upon the modal step of their *empirical continuum.*

Uniformity of behavior (where there is an equal probability for the phenomenon to either occur or fail to occur) is indicated by the kurtosis. To return to our analogy in the binomial expansion, it is the case where, in the formula $(q + p)^s$, s is small and the possibilities of response limited. From formula (8) it follows that as $s \rightarrow \infty$ the first two expressions on the right-hand side of the equation approach zero and α_4 approaches 3.

When α_4 of a distribution is significantly[†] greater than 3 we may say that the response in a given situation is more uniform than could be explained by chance.

The Application of the Method of Moments

To compare the methods outlined above with those of the "*J*-curve

[*] The variance of α_3 is given by Fisher, (6, p. 79) as $\dfrac{(6N)\,(N-1)}{(N-2)\,(N+1)\,(N+3)}$.

To be significantly different from zero, α_3 must (in the case of large samples) be 2.58 or more times greater than the square root of its standard error.

[†] σ^2 of α_4 is given by Fisher (6) as $\dfrac{24N(N-1)^2}{(N-3)\,(N-2)\,(N+3)\,(N+s)}$.

hypothesis," the data presented by Allport and Solomon (3) on length of conversation in church, library, and club-room have been reanalyzed. The results are presented in Tables 3 and 4.

The method of moments reveals that both conformity and uniformity beyond that which might be attributed to chance factors alone exist in all three situations (See columns 2 and 4, Table 3). This fails to confirm the finding of Allport and Solomon to the effect that conformity did not exist in the club-room situation. Table 4 indicates that any difference in skewness between the clubroom and library situation can be explained on the basis of chance alone. It is impossible to maintain, as Allport and Solomon do, that conformity exists in one and not the other of these situations.

Summary and Conclusions

The "*J*-curve hypothesis" of conforming behavior was examined and found to present three main weaknesses: (1) the necessarily *post hoc* definition of conformity in any given situation; (2) the failure of the advocates of the hypothesis to secure equal "telic" units; and (3) the failure of the hypothesis to provide any means of comparing the difference in conformity between two situations. A mathematical basis for the measurement of conformity was briefly out-

TABLE 3

A Comparison of α_3 and α_4 and Their Standard Errors in the Three Situations Presented by Allport and Solomon

Situation	N	α_3	σ_{α_3}	α_4	σ_{α_4}
Church	200	2.50	.1720	11.15	.3417
Library	802	1.96	.0860	8.94	.1726
Clubroom	400	1.78	.1183	8.15	.2420

TABLE 4

The Differences and the Significance of the Differences of α_3 and α_4 of Allport and Solomon Data

Situation	Diff._{α_3}	$\sigma_{\text{Diff.}_{\alpha_3}}$	t_{α_3}	Diff._{α_4}	$\sigma_{\text{Diff.}_{\alpha_4}}$	t_{α_4}
Church-Library	.54	.1923	2.81	2.21	.3827	>3
Church-Clubroom	.71	.2088	3.40	3.00	.4192	>3
Library-Clubroom	.18	.1463	1.23	.79	.2980	2.65

lined and the method of moments suggested as a more satisfactory way of dealing with measurements of behavior in this field. Finally, data presented by Allport and Solomon were subjected to re-analysis by the method of moments and the results presented.

From what has been presented here we may conclude:

(1) that the "*J*-curve hypothesis" is inadequate for the purposes for which it was designed.

(2) that the "*J*-curve hypothesis" creates in conformity differences which have been shown to be statistically insignificant. (Library vs. club-room situations).

(3) that no special technique is required for the analysis of conformity data.

(4) that the method of moments is adequate to give quantitative expression to conformity and uniformity of behavior, and

(5) that this method distinguishes between conformity (skewness) and uniformity (kurtosis).

REFERENCES

1. Allport, F. H. The *J*-curve hypothesis of conforming behavior. *J. soc. Psychol.*, 1934, 5, 141-183.
2. Allport, F. H. Rule and custom as individual variation of behavior distribution along a continuum of conformity. *Amer. J. Sociol.*, 1939, 44, 897-921.
3. Allport, F. H., and Solomon, R. S. Lengths of conversation; A conformity situation analyzed by the telic continuum and *J*-curve hypothesis. *J. abn. soc. Psychol.*, 1939, 34, p. 419.
4. Dickens, M., and Solomon, R. S. The *J*-curve hypothesis—Certain aspects clarified. *Sociometry*, 1938, 1, p. 277.
5. Dudycha, G. J. An examination of the *J*-curve hypothesis based on punctuality distribution. *Sociometry*, 1937, 1, 144-154.
6. Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver & Boyd, 1936.
7. Pearson, K., and Lee, A. On the law of inheritance of man. *Biometrika*, 1903, 2, 351-462.
8. Solomon, R. S. An index of conformity based on the *J*-curve hypothesis. *Sociometry*, 1939, 2, 63.

PSYCHOMETRIKA—VOL. 7, NO. 1
MARCH, 1942

THE SCALING OF TEST SCORES BY THE METHOD OF PAIRED COMPARISONS*

LOUISE T. GROSSNICKLE
THE UNIVERSITY OF CHICAGO

The purpose of this study is to investigate, by the method of paired comparisons, a possible scaling of individuals who have made certain test scores, such that the additive property will be satisfied and such that a stability in scaling will be maintained ,—in other words, a scaling such that the scaled score of an individual will remain relatively the same regardless of the grouping of individuals in which he may be placed. The results show that it is possible to utilize psychophysical methods in psychological and educational test situations. Among the major findings are that Case V of the Law of Comparative Judgment is applicable to the data in this problem, the method of dividing the intermediate category equally between the greater and the less was the best of three possible methods, internal consistency was satisfied, and, finally, when a new test of stability was applied, it was found that the distances between the hypothetical individuals remain the same.

It is well known that the raw scores of any test fail to satisfy the additive property and are merely expressed in terms of the relative numbers of items passed. With this in mind, the purpose of the study here reported is to investigate, by the method of paired comparisons, a possible scaling of individuals who have made certain test scores, such that the additive property will be satisfied and such that a stability in scaling will be maintained, in other words a scaling such that the distance between the scaled scores of any two individuals will remain the same regardless of the grouping of individuals in which they may be placed. In this particular problem, Thurstone's Law of Comparative Judgment (5) will be subjected to various tests to see whether its reliability is consistently maintained.

Psychophysical methods usually have been applied in the fields of sensory discrimination and in scaling attitudes and opinions. The application here of the method of paired comparisons lies in a different field, that of the mental and the educational test, wherein the individuals are the stimuli and are scaled accordingly.

For this purpose, a homogeneous vocabulary test, consisting of one hundred items in multiple-choice form, was selected from a Biological Science examination developed at The University of Chicago.

* The writer wishes to express appreciation for the invaluable help and guidance of Professor Harold O. Gulliksen and also to Professor Marion W. Richardson who suggested the problem and made valuable suggestions.

From the group of college freshmen who took this examination, one hundred were selected by alphabetical order of last names. The test papers were made available by the Board of Examinations. If any particular person omitted fifty per cent of the items, his test paper was discarded. This criterion necessitated the elimination of only twelve tests from among those examined.

Table I is an illustration showing how the score matrix for the hundred persons and the hundred items was set up. An unsuccessful response is designated by an x, an omitted item by an o, a right an-

TABLE 1

Record of Test Items Answered Successfully or Unsuccessfully by the Hundred Persons Taking the Test
(An x indicates a wrong answer, an o an omitted item, a blank space a correct answer)

Hypothetical Individuals	Persons	Items												Total Score
		1	2	3	4	5	6	7	8	9	10	. . . .	100	
1	1				x									93
	2			x							x			93
	3	x									o		x	86
	4		x	x		x	x							84
	5	x	x	x				x					x	83
2	6			x	x								o	82
	7			x				x	x				x	81
	8	x	x		x				o				x	80
	9			x	x		x				x		x	79
	10		x										x	79
3	11						x			x	x		o	78
	12						x	x	x		x		o	78
	13						x	x	x		x		x	77
	14	x		x	x	x			x					76
	15		x	x									x	75
•	•													•
•	•													•
•	•													•
20	96	x	x	x			x		x				o	37
	97	o	x	x	x	x			x		x		o	37
	98	x		x	o		o	o	x	x	o		o	34
	99	x	x	x	x		x	x		x	x		o	33
	100	o	o	o	x		x		x	x	x		o	29

swer by a blank space. The hundred persons were originally arranged in rank order according to total scores. Regarding the table horizontally, the response of each person to each item of the test may be observed. Viewing Table 1 vertically, each item may be said to "judge" the persons taking the test in reference to success or failure. The hundred persons, arranged in rank order, were then grouped by fives, forming twenty hypothetical individuals. Thus the hypothetical

TABLE 2

The Number of Times the Hypothetical Individual Given at the Top of the Column was More, Less, or Equally Successful when Compared with Those at the Left (The top number for each individual indicates greater, the middle number less, the lower number equal success)

Hypothetical Individuals	Hypothetical Individuals																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2	34 15 51																			
3	46 15 39	34 22 44																		
4	51 12 37	37 13 50	29 23 48																	
5	49 13 38	39 15 46	34 23 43	28 24 48																
6	54 12 34	38 14 48	36 22 42	33 23 44	30 24 46															
7	53 09 38	44 10 46	42 22 36	37 24 39	33 24 43	32 27 41														
8	60 09 31	55 12 33	42 16 42	48 25 27	37 20 43	40 19 41	33 27 40													
9	65 08 27	56 08 36	46 10 44	48 24 28	45 22 33	43 19 38	41 26 33	33 25 42												

TABLE 2 (continued)

The Number of Times the Hypothetical Individual Given at the Top of the Column was More, Less, or Equally Successful when Compared with Those at the Left (The top number for each individual indicates greater, the middle number less, the lower number equal success)

Hypothetical Individuals	Hypothetical Individuals																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10	67	58	54	47	45	43	43	37	35											
	04	09	14	18	21	20	25	27	32											
	29	33	32	35	34	37	32	36	33											
11	66	59	57	48	41	46	45	39	38	30										
	07	09	12	19	18	19	23	28	31	30										
	27	32	31	33	41	35	32	33	31	40										
12	69	67	62	59	50	54	46	40	36	36	36									
	16	02	15	16	16	16	18	20	24	27	29									
	15	31	23	25	34	30	36	40	40	37	35									
13	73	60	56	52	48	45	43	44	36	41	37	34								
	06	06	13	11	11	16	14	26	30	25	34	35								
	21	34	31	37	41	39	43	30	34	34	29	31								
14	73	68	63	60	54	51	52	44	42	42	37	34	35							
	03	04	10	14	15	14	20	22	23	28	25	29	30							
	24	28	27	26	31	35	28	34	35	30	38	37	35							
15	71	67	63	60	53	48	48	47	45	47	44	40	38	38						
	05	08	07	11	11	14	15	24	21	21	26	29	24	34						
	24	25	30	29	36	38	37	29	34	32	30	31	38	28						
16	82	72	71	66	65	62	58	55	48	54	45	46	51	39	35					
	03	04	09	13	13	10	11	15	19	21	23	27	26	31	29					
	15	24	20	21	22	28	31	30	33	25	32	27	23	30	36					
17	78	71	69	63	58	58	59	57	54	43	51	46	45	43	37	39				
	04	02	06	12	09	14	10	18	18	17	18	20	20	24	29	32				
	18	27	25	25	33	28	31	25	28	40	31	34	35	33	34	29				
18	83	84	73	70	69	70	67	60	57	55	58	53	51	50	45	38	36			
	03	03	04	08	06	07	09	09	10	17	17	19	15	20	28	25	34			
	14	13	23	22	25	23	24	31	33	28	25	28	34	30	27	37	30			
19	83	82	81	77	74	68	68	65	60	59	60	59	60	51	55	48	41	43		
	04	02	02	11	06	06	07	15	12	15	13	11	14	16	21	28	23	27		
	13	16	17	12	20	26	25	20	28	26	27	30	26	33	24	24	36	30		
20	84	87	85	84	80	80	80	80	75	72	65	69	72	66	55	57	54	57	45	
	01	01	02	03	04	04	03	05	09	04	10	06	08	12	10	17	17	15	19	
	15	12	13	13	16	16	17	15	16	24	25	25	20	22	35	26	29	28	36	

individual of top rank, for example, comprises the five highest ranking persons, and so on until the twentieth hypothetical individual of lowest rank includes the five persons of lowest rank. The reasons for grouping the one hundred persons into these twenty hypothetical individuals were to minimize the labor involved in the method of paired comparisons and to insure greater stability in the judgments of the items. The judgments of each of the items on the twenty hypothetical individuals were obtained. For example, in Table 1, item one judges hypothetical individual two (where four persons out of five were successful) greater than individual one (with three persons answering the item correctly). Similarly, in the case of item two, individuals one and two are judged equal, for two persons were successful within each. Comparing individuals one and three in regard to item ten, individual one (with three persons successful) is judged greater than individual three (where only two persons were successful in answering the item). An omitted item, as well as a wrong answer, was counted as unsuccessful.

Then each hypothetical individual was used as a standard and compared with every other one in regard to success on each item, and the frequencies of greater, less, and equal tabulated in Table 2. For example, individual one has been judged greater than individual two on thirty-four items out of the hundred, individual one is judged less than two on fifteen items, and they are judged equally successful on fifty-one items. As in the field of mental test theory, the items "test" or judge the group of persons, while in the usual psychophysical problem the reverse is true, the judges evaluating the stimuli or items. In Table 3, the frequencies translated into the corresponding proportions are given. The proportion of each as compared with itself was assumed to be .50 and the number of terms in the psychophysical table $n(n-1)/2$.

In this problem there are three possible ways of dealing with the intermediate category: dividing it proportionally to the greater and less categories, dividing it equally between them, and dividing it so as to split the base line of the equal category. These three methods may be described in more detail as follows:

1. *Dividing the equal category proportionally to the greater and the less categories.*

   Where G means the number in the greater category, L is the number in the less category, and E the number in the equal category;

   $G + \frac{G}{G+L} E =$ Value of the greater category when

TABLE 3

Proportions of the Times the Hypothetical Individual Given at the Top was More Successful than Those at the Left*
(The method of deriving the proportions was to assign one-half the equal category to the greater and one-half to the less category)

Hypo-thetical Indi-viduals	Hypothetical Individuals																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	500	405	345	305	320	290	280	245	215	185	205	235	165	150	170	105	130	100	105	085
2	595	500	440	380	380	380	330	285	260	255	250	175	230	180	205	160	155	095	100	070
3	655	560	500	470	445	430	400	370	320	300	275	265	285	235	220	190	185	155	105	085
4	695	620	530	500	480	450	435	385	380	355	355	285	295	270	255	235	245	190	170	095
5	680	620	555	520	500	470	455	415	385	380	385	330	315	305	290	240	255	185	160	120
6	710	620	570	550	530	500	475	395	380	385	365	310	355	315	330	240	280	185	190	120
7	720	670	600	565	545	525	500	470	425	410	390	360	355	340	335	265	255	210	195	115
8	755	715	630	615	585	605	530	500	460	450	445	400	410	390	385	300	305	245	250	125
9	785	740	680	620	615	620	575	540	500	485	465	440	470	405	380	355	320	265	260	170
10	815	745	700	645	620	615	590	550	515	500	500	455	420	430	370	335	370	310	280	160
11	795	750	725	645	615	635	610	555	535	500	500	465	485	440	410	390	335	295	265	225
12	765	825	735	715	670	690	640	600	560	545	535	500	505	475	445	405	370	330	260	185
13	835	770	715	705	685	645	645	590	530	580	515	495	500	475	430	375	375	320	270	180
14	850	820	765	730	695	685	660	610	595	570	560	525	525	500	480	460	405	350	325	230
15	830	795	780	745	710	670	665	615	620	630	590	555	570	520	500	470	460	415	330	275
16	895	840	810	765	760	760	735	700	645	665	610	595	625	540	530	500	465	435	400	300
17	870	845	815	755	745	720	745	695	680	630	665	630	625	595	540	535	500	490	410	315
18	900	905	845	810	815	815	790	755	735	690	705	670	680	650	585	565	510	500	420	290
19	895	900	895	830	840	810	805	750	740	720	735	740	730	675	670	600	590	580	500	370
20	915	930	915	905	880	880	885	875	830	840	775	815	820	770	725	700	685	710	630	500

* All entries have been multiplied by 1,000 to eliminate the decimal.

that amount of the equal category proportional to it has been added;

$L + \frac{L}{G+L}E =$ Value of the less category with that amount of the equal category proportional to it added;

$$\frac{G + \frac{G}{G+L}E}{G + \frac{G}{G+L}E + L + \frac{L}{G+L}E} = \frac{G}{G+L}. \qquad (1)$$

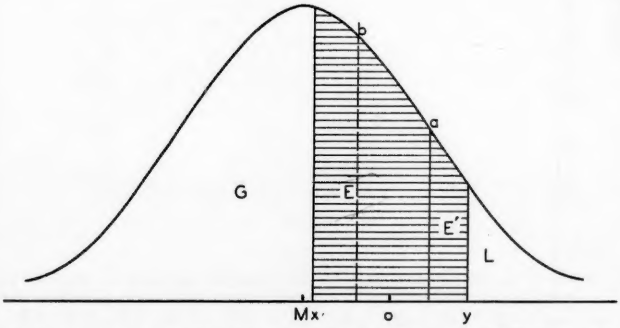
Dividing proportionally, then, gives the same results as disregarding the equal category entirely and considering the ratio of the greater divided by the greater plus the less categories.

2. *Dividing the equal category equally between the other two categories.*

This proportion takes the form:

$$\frac{G + \frac{1}{2}E}{100}. \qquad (2)$$

This method gives a smaller value to the greater category than the preceding method where the greater category received the greater part of the equal category. This may be illustrated graphically.



The shaded area represents the equal category divided in such a way as to give the proportion E to the greater category and the proportion E' to the less.

$$E : E' = G : L. \qquad (3)$$

On the other hand, if the equal category is divided equally between

G and L, the division indicated by line a will have to move toward the mean of the distribution (dotted line b). Accordingly, the proportions and their corresponding sigma values will be smaller than in the case of dividing proportionally.

3. *Splitting the base line of the equal category.*

   In this case, the sigma value corresponding to $P = \frac{G}{100}$ is added to the sigma value corresponding to $P = \frac{G + E}{100}$ and the two are averaged.

Referring to the figure illustrated above, the base line xy of the shaded area is to be equally divided such that $xo = oy$. While dividing equally makes the two parts of the intermediate category equal in terms of proportion or area of the normal probability curve, dividing by splitting the base line of the equal category makes them equal in terms of the corresponding sigma values. This latter method will

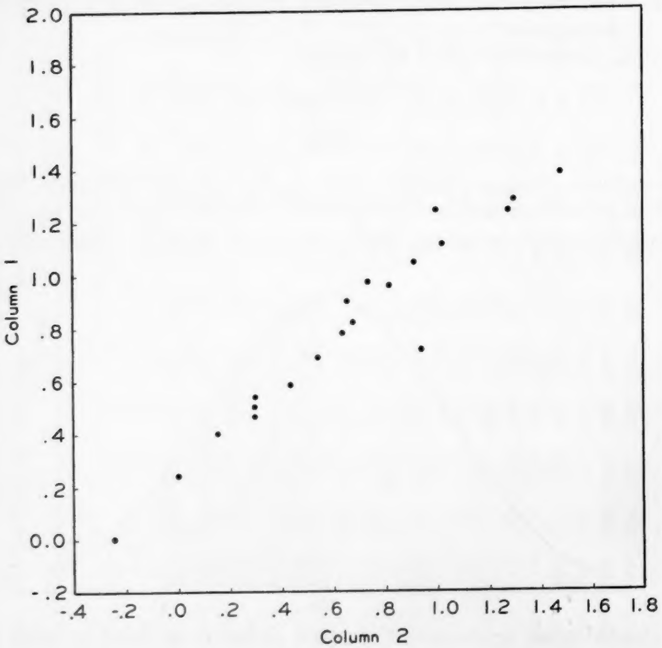


FIGURE 1
Sigma Values of Column One Plotted Against Those of Column Two Using the Proportion Obtained by Dividing Equally

give the larger sigma value to the greater category than in the case of dividing equally.

Two tests were applied to determine which of the three types of proportion is best adapted to this problem.

1. (Test of linearity by inspection—Figures 1-4)

As a rough check, the trend of linearity was determined by inspection. Using sigma values corresponding to the type of proportions dividing equally, sixty-two graphs were plotted. Each column of sigma values, corresponding to the x_{ba} values of the Law of Comparative Judgment, was plotted in turn against each of the three following columns. Column number twenty was plotted against columns one, two, and three respectively. Figures 1 and 2 are samples chosen to demonstrate the linearity obtained using the type of proportion secured by dividing equally, as superior to that of the linearity resulting from the other two methods to be discussed. Similarly, twenty-seven graphs were constructed for the sigma values of the proportion

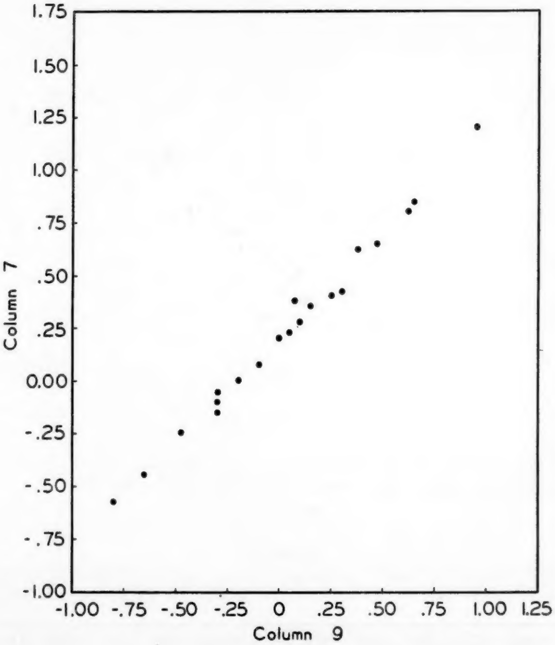


FIGURE 2
Sigma Values of Column Seven Plotted Against Those of Column Nine Using the Proportion Obtained by Dividing Equally

dividing proportionally, using in this case only immediately adjacent columns and a few other cases in which linearity was noticeably good or relatively poorer in the case of the previous graph. (As an example, see Figure 3.) Comparing the two sets of graphs, a majority of cases resulted in which the proportion obtained by dividing equally gave plots which were definitely more linear than dividing proportionally. In a smaller number of cases the two proved about equal. It was found by inspection also that the slopes of these plots closely approximated unity. This finding will be used later in verifying the ap-

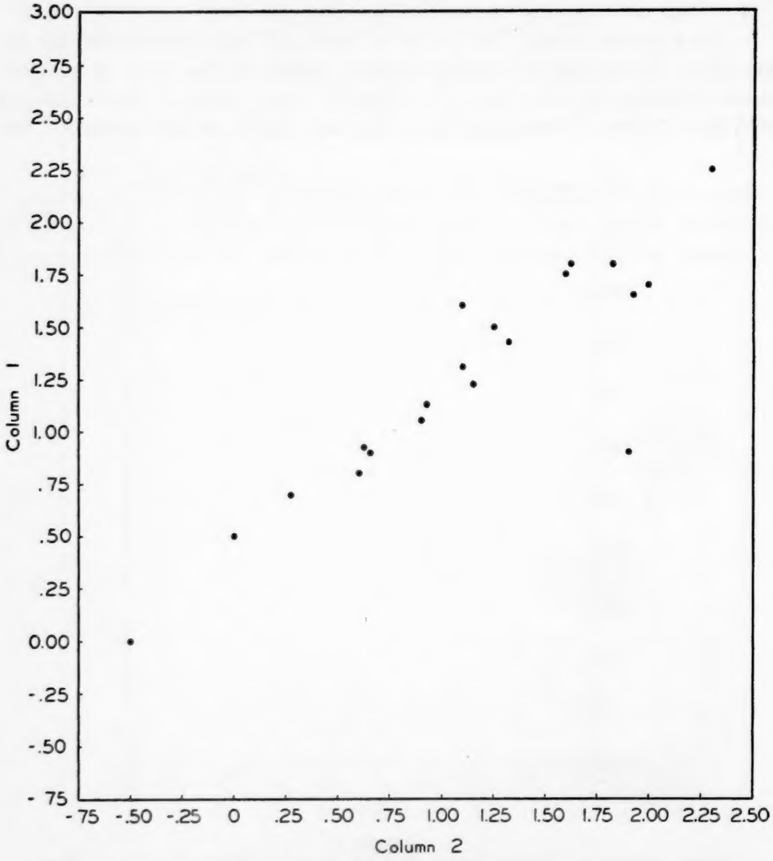


FIGURE 3
Sigma Values of Column One Plotted Against Those of Column Two Using the Proportion Obtained by Dividing Proportionally

plicability of Case V of the Law of Comparative Judgment to these data. Using some cases in which linearity was relatively superior or inferior in the case of the above-mentioned proportions, and others representing about equal linearity, eight graphs were plotted, using sigma values for the proportion obtained by splitting the base line of the equal category. All plots obtained by this latter method were less linear than in the cases of the other two methods used. Here the plot tended to be curvilinear rather than linear (Figure 4). The method of dealing with the intermediate category by splitting the base line of the equal category was accordingly dropped as not worth while for further investigation.

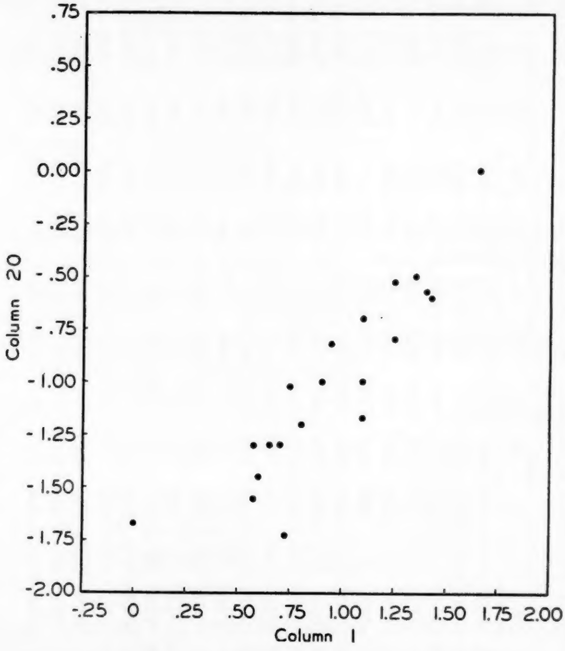


FIGURE 4
Sigma Values of Column Twenty Plotted Against Those of Column One Using the Proportion Obtained by Splitting the Base Line of the Equal Category

2. Discrepancy between variance and covariance—

A more rigorous test applied in finding the best proportion for this particular problem, using the method of least squares, was to find the discrepancy between the variance and twice the covariance of any two columns of sigma values. Finding the discrepancy of the vari-

TABLE 4

Scale Separations in Terms of Standard Deviations Corresponding to Proportions in Table 3*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	000	–240	–399	–510	–468	–553	–583	–690	–789	–897	–824	–723	–974	–1036	–954	–1254	–1126	–1282	–1254	–1372
2	240	000	–151	–306	–306	–306	–440	–568	–643	–659	–675	–935	–739	– 915	–824	– 995	–1015	–1311	–1282	–1476
3	399	151	000	–075	–138	–176	–253	–332	–468	–524	–598	–628	–568	– 723	–772	– 878	– 897	–1015	–1254	–1372
4	510	306	075	000	–050	–126	–164	–292	–306	–372	–372	–568	–539	– 613	–659	– 723	– 690	– 878	– 954	–1311
5	468	306	138	050	000	–075	–113	–215	–292	–306	–292	–440	–482	– 510	–553	– 706	– 659	– 897	– 995	–1175
6	553	306	176	126	075	000	–063	–266	–306	–292	–345	–496	–372	– 482	–440	– 706	– 583	– 897	– 878	–1175
7	583	440	253	164	113	063	000	–075	–189	–228	–279	–359	– 372	– 413	–426	– 628	– 659	– 806	– 860	–1200
8	690	568	332	292	215	266	075	000	–100	–126	–138	–253	–228	– 279	–292	– 524	– 510	– 690	– 675	–1150
9	789	643	468	306	292	306	189	100	000	–038	–088	–151	–075	– 240	–306	– 372	– 468	– 628	– 643	– 954
10	897	659	524	372	306	292	228	126	038	000	000	–113	–202	– 176	–332	– 426	– 332	– 496	– 583	– 995
11	824	675	598	372	292	345	279	138	088	000	000	–088	–038	– 151	–228	– 279	– 426	– 539	– 628	– 755
12	723	935	628	568	440	496	359	253	151	113	088	000	013	– 063	–138	– 240	– 332	– 440	– 643	– 897
13	974	739	568	539	482	372	372	228	075	202	038	–013	000	– 063	–176	– 319	– 319	– 468	– 613	– 915
14	1036	915	723	613	510	482	413	279	240	176	151	063	063	000	–050	– 100	– 240	– 385	– 454	– 739
15	954	824	772	659	553	440	426	292	306	332	228	138	176	050	000	– 075	– 100	– 215	– 440	– 598
16	1254	995	878	723	706	706	628	524	372	426	279	240	319	100	075	000	088	– 164	– 253	– 524
17	1126	1015	897	690	659	583	659	510	468	332	426	332	319	240	100	088	000	– 025	– 228	– 482
18	1282	1311	1015	878	897	897	806	690	628	496	539	440	468	385	215	164	025	000	– 202	– 553
19	1254	1282	1254	954	995	878	860	675	643	583	628	643	613	454	440	253	228	202	000	– 332
20	1372	1476	1372	1311	1175	1175	1200	1150	954	995	755	897	915	739	598	524	482	553	332	000

* All entries have been multiplied by 1,000 to eliminate the decimal.

ance and twice the covariance should be a better test than finding the correlation coefficient, where the formula $y = ax + b$ holds, because in this particular problem the slope of any two columns plotted is a slope of unity and the value of a in the equation is one. Then, using x and y as deviations from the mean of each column of sigma values,

$$y = x + b\,. \tag{4}$$

But $b = M_y - M_x = 0$; and using the new origin,

$$\sum (y - x)^2 = 0\,. \tag{5}$$

Expanding and transposing,

$$\sum x^2 + \sum y^2 = 2 \sum xy\,. \tag{6}$$

Results in comparing the proportions dividing proportionally and dividing equally, using sigma values, show the following discrepancies: (The columns were chosen which seemed by inspection of the plots to be about equal in linearity for each of the two types of proportions. Two samples from the middle of the range of individuals, one at the end, and one at the beginning, are used.)

	Discrepancy	
	Dividing Equally	Dividing Proportionally
Columns 4 and 3 compared	.0893	.6700
Columns 11 and 12 compared	.1538	1.0239
Columns 10 and 11 compared	.1262	.4490
Columns 16 and 19 compared	.0279	.7873

In this more rigorous test than the method of inspection, stated above, again, dividing equally showed the less discrepancy and proved the better proportion to use in this particular problem. Accordingly, the equal category was assigned, fifty per cent to the greater and fifty per cent to the less category, and the final scale values were therefore determined on the basis of this proportion. The sigma values resulting from this procedure are presented in Table 4.

Thurstone's Law of Comparative Judgment (5) may be stated as

$$S_b - S_a = X_{ba}\sqrt{\sigma_b{}^2 + \sigma_a{}^2 - 2r_{ab}\,\sigma_a\sigma_b}\,. \tag{7}$$

In this study, Case V, the simplest form of the Law of Comparative Judgment, was tested for applicability. This form, in addition to the assumptions of normality of discriminal processes, of approximate equality of the discriminal dispersions, and of zero correlations between judgments of any two stimuli, when applied to a group, assumes that the discriminal dispersions are equal. Formula (7) then becomes

$$S_b - S_a = X_{ba} \sqrt{2\sigma^2} \,, \tag{8}$$

where S_b is the scale value of stimulus 1 (hypothetical individual 1).

S_a is the scale value of hypothetical individual 2.

X_{ba} is the sigma value corresponding to the observed proportion of judgments $b > a$.

σ^2 refers to the discriminal dispersions (variability) assumed equal for all stimuli (individuals) in this problem.

In assuming Case V for the solution of this problem, we must provide some sort of experimental check for two assumptions: first, normality of the distribution of discriminal processes (of the variability of the individuals) and second, equality of the discriminal dispersions.

Professor Thurstone (9) has suggested that a check for these two assumptions may be graphically determined by plotting the sigma values of proportions $b > a$ of any two adjacent columns. If the plot is linear, the assumed normal distribution of discriminal processes is correct; if there is a slope of unity, the discriminal dispersions are equal. This check has already been utilized when these sigma values were plotted against each other in selecting one of the three ways of dealing with the intermediate category. This means that in this problem, since there was linearity and unity of slope, the variability of the hypothetical individuals is equal throughout and the distributions are normal, and therefore Case V may be used.

Having satisfied the assumptions of Case V, the solution of the problem is continued with this case.

Using the subscript k to signify each stimulus in turn compared with the standard,

$$S_b - S_k = X_{bk} \sqrt{2}\, \sigma \,; \tag{9}$$

$$S_a - S_k = X_{ak} \sqrt{2}\, \sigma \,. \tag{10}$$

Subtracting and summing for all values of k, dividing by n, and letting the sigma value be the unit of the scale,

$$S_b - S_a = \frac{\sum (X_{bk} - X_{ak})}{n} \sqrt{2}. \tag{11}$$

TABLE 5

Differences Between Each S Value and Its Adjacent Value and the Unweighted Means of the Differences*

R	$d_{1,2}$	$d_{2,3}$	$d_{3,4}$	$d_{4,5}$	$d_{5,6}$	$d_{6,7}$	$d_{7,8}$	$d_{8,9}$	$d_{9,10}$	$d_{10,11}$	$d_{11,12}$	$d_{12,13}$	$d_{13,14}$	$d_{14,15}$	$d_{15,16}$	$d_{16,17}$	$d_{17,18}$	$d_{18,19}$	$d_{19,20}$
1	......	159	111	–042	085	030	107	099	108	–073	–101	251	062	–082	300	–128	156	–028	118
2	240	......	155	000	000	134	128	075	016	016	260	–196	176	–091	171	020	296	–029	194
3	248	151	......	063	038	077	079	136	056	074	030	–060	155	049	106	019	118	239	118
4	204	231	075	......	076	038	128	014	066	000	196	–029	074	046	064	–033	188	076	357
5	162	168	088	050	......	038	102	077	014	–014	148	042	028	043	153	–047	238	098	180
6	247	130	050	051	075	......	203	040	–014	053	151	–124	110	–042	266	–123	314	–019	297
7	143	187	089	051	050	063	......	114	039	051	080	013	041	013	202	031	147	054	340
8	122	236	040	077	–051	191	075	......	026	012	115	–025	051	013	232	–014	180	–015	475
9	146	175	162	014	–014	117	089	100	......	050	063	–076	165	066	066	096	160	015	311
10	238	135	152	066	014	064	102	088	038	......	113	089	–026	156	094	–094	164	087	412
11	149	077	226	080	–053	066	141	050	088	000	......	–050	113	077	051	147	113	089	127
12	–212	307	060	128	–056	137	106	102	038	025	088	......	068	075	102	092	108	203	254
13	235	171	029	057	110	000	144	153	–127	164	043	–005	......	113	143	000	149	145	302
14	121	192	110	103	028	069	134	039	064	025	088	000	063	......	050	140	145	069	285
15	130	052	113	106	113	014	134	–014	–026	104	090	–038	126	050	......	025	115	225	158
16	259	117	155	017	000	078	104	152	–054	147	039	–079	219	025	075	......	076	089	271
17	111	118	207	031	076	–076	149	042	136	–094	094	013	079	140	012	088	......	203	254
18	–029	296	137	–019	000	091	116	062	132	–043	099	–028	083	170	051	139	025	......	351
19	–028	028	300	–041	117	018	185	032	060	–045	–015	030	159	014	187	025	026	202	......
20	–104	104	061	136	000	–025	050	196	–041	240	–142	–018	176	141	074	042	–071	221	332
d	2382	3034	2320	928	608	1124	2276	1557	619	692	1439	–290	1922	976	2399	425	2647	1924	5136
M_d	125	160	122	049	032	059	120	082	033	036	076	–015	101	051	126	022	139	101	270
$\sqrt{2}M_d$	177	226	173	069	045	083	170	116	047	051	107	–021	143	072	178	031	197	143	382

* All entries have been multiplied by 1,000 to eliminate the decimal.

In applying formula (11) to the problem, the quantity $(X_{bk} - X_{ak})$ for each value of k is entered in a table of differences for each two adjacent individuals as stimuli; that is, $d_{1,2}$, $d_{2,3}$ as given in Table 5. The mean of each column of differences is calculated as a better measure than any one of them, and each mean difference is multiplied by $\sqrt{2}$ as Case 5 requires. This value $M_{\text{diff.}}\sqrt{2}$ gives the final scale separations in terms of the standard deviation of the discriminal dispersions (variability of the individuals). Then the scale is built up, a value of zero being assigned to the lowest scale value and these values being accumulated (Table 6).

TABLE 6

Unweighted and Weighted Scale Values

Unweighted Scale Values	Weighted Scale Values	Unweighted Scale Values	Weighted Scale Values
2.389	2.299	1.232	1.179
2.212	2.101	1.125	1.074
1.986	1.885	1.146	1.095
1.813	1.727	1.003	.961
1.744	1.662	.931	.887
1.699	1.620	.753	.724
1.616	1.537	.722	.686
1.446	1.376	.525	.509
1.330	1.269	.382	.365
1.283	1.227	.000	.000

The question arises as to whether the use of weighted values might differ significantly from the unweighted. An adaptation of the Müller-Urban weights (2) was used, similar in principle to Thurstone's weighting formula (10), which weights inversely to the square of the standard error, or in proportion to the reliability of the original value.

The weight of a difference d_{ba} is given by the formula

$$W_{ba} = \frac{1}{\dfrac{1}{W_{ak}} + \dfrac{1}{W_{bk}}}, \tag{12}$$

in which W_{ak} and W_{bk} are the Müller-Urban weights corresponding to the proportions P_{ak} and P_{bk}, respectively, k meaning any other stimulus. Multiplying each difference down the column by its appropriate weight, the weighted differences are found, and their mean difference calculated and multiplied by $\sqrt{2}$.

Now Case V becomes

$$S_{Wb} - S_{Wa} = X_{Wba} \sqrt{2} \qquad (13)$$

Then the scale is built up as in the case of the unweighted values (Table 6).

In testing for the significance of the difference between the means of weighted and unweighted scale values, several procedures have been considered. A test of significance directly comparing the two sets of scale values (weighted against unweighted) was not possible because the arbitrary origin of each scale affects the numerator of the resulting critical ratio and because of the accumulation of differences as the number of scale values which one happens to have increases. In view of this, the method used in this study is that of finding the significance of the difference of the mean differences of the respective columns before they are accumulated into scale values. Fisher's method of handling the differences directly was used. This method is applicable when the variables are correlated, and corrects for a small sample. The critical ratio is 2.2, being significant only at the five per cent level. Thus the results are not significant under the crucial tests at the one and two per cent levels. On this basis, the weighted scale was discarded as not worth the trouble of weighting, and the unweighted scale values adopted. Since there were no proportions $P_{a>b}$ above .97 or below .03 in using the proportion dividing equally, all proportions in the table were retained and given equal weight.

For the proof of internal consistency of the scale, there are three checks presented in this problem, one based on an inspection of the columns of differences, one based on the mean of the discrepancies between the actual and the theoretical proportions, and a test for stability.

Professor Thurstone has suggested a quick check for proving internal consistency by inspecting each column of differences derived from sigma values (Table 5). These x distances should hang together; that is, there should be no systematic drifts in values up or down the column. On inspecting Table 5, we see that no column gives any systematic increase or decrease in values. However, this is only a rough check.

The real proof for internal consistency which Professor Thurstone has provided (8) is that based on the algebraic mean of the discrepancies and the average discrepancy, disregarding sign between actual and theoretical proportions. Starting with the final scale values and working backward to the theoretical proportions demanded of them, the differences between these values and those of the original proportions were obtained and the mean of these discrepancies calculated to be .00018. The distribution of the discrepancies has been plotted and the results shown in Figure 5. The distribution proves to be symmetrical. The average discrepancy, disregarding signs, is .014.

Another check for the reliability of the scale which can be applied in this particular problem is the test of stability. Stability in this situation means that the scale distance between any two individuals taking a test remains unchanged no matter in what new group of individuals they are placed. In this test, four hypothetical individuals were chosen at random from among the twenty hypothetical individ-

TABLE 7

Test Items Answered by Thirty Additional Cases and Twenty Original Persons in the New Data for the Test of Stability

(The hypothetical individuals 1, 3, 8, 9 were taken at random from the original scale. Their ranks in the original scale were 1, 6, 17, and 18 respectively)

Hypothetical Individuals	Persons	Items												Total Score
		1	2	3	4	5	6	7	8	9	10	....	100	
1	1				x									93
	2			x							x			93
	3	x									o		x	86
	4		x	x		x	x							84
	5	x	x	x				x					x	83
2	6	x												89
	7			x							x			83
	8		x	x				x					o	83
	9						o		o		x		o	74
	10				x		x						x	73
.	.													.
.	.													.
.	.													.
9	41	x	x	x	o		o	o	o		o			47
	42	x	x	x		x	o	o	x		o		x	46
	43	x	x	x	x		x		x	x	x		x	46
	44			x	x	x			x		x		x	46
	45		x		x		x			x			o	44
10	46		x		x	x	x		x		x		x	49
	47		x	x	x		o		x		o		o	48
	48	o	x	x		x	x		o				o	44
	49	x	x	x	o	x		x	x	x	o		o	38
	50	x	x		x		x	o	x	x	x		o	31

uals originally scaled, to be placed along with six new hypothetical individuals forming a new scale. For these six new individuals, the test records of thirty new persons who took this same test were selected at the Board of Examinations in the same manner as those originally chosen, that is, by alphabetical order of last names. These were again ranked in terms of final scores and grouped by fives, forming six hypothetical individuals. Then the four hypothetical individuals of the old scale were slipped in among them, all being ranked according to the rank of the raw scores, as is shown in Table 7. Then the new scale was built up, using the four individuals from the original scale and the six from the new data.

Since the origin of the old and the new populations comprising this new scale are arbitrary, one can not expect the scaled scores of these four individuals to be the same in the new scale as they were in the old one. However, if the scale is stable, the distance between them will be the same in both scales. Results clearly indicate, as is shown in Table 8 and Figure 6, that the scale distance between any two of the individuals common to both scales is the same on the new scale as in the old, and thus that the scale has a high degree of stability.

TABLE 8

Record of Distances Between Any Two Hypothetical Individuals Common to Both Old and New Scales in the Test for Stability
Hypothetical Individuals Common to Both Scales

Rank in Old Scale	Rank in New Scale
1	1
6	3
17	8
18	9

The distance between any two of the four hypothetical individuals should remain the same if the test for stability holds. (The mean differences of both scales are used in the comparison.)

M_d1-3 should equal M_d 1- 6 . . . . .4910 = .4880

M_d3-8 should equal M_d 6-17 . . . . .6689 = .6915

M_d8,9 should equal M_d17,18 . . . . .1396 = .1393

The scaled scores for the hypothetical individuals, in both the old and new sets of data, were plotted against their averaged raw scores, respectively, and the plots found to be fairly linear. The question may be raised as to what the use of building up a scale may be, if plots of the raw and the scaled scores tend toward linearity.

However, there are basic differences between the scaled scores and the raw scores. The scaled scores have fulfilled the mathematical additive function. Therefore, they are based on a rational function. On the other hand, the raw scores are not. It may be pointed out that the linearity for the data of this study may be attributable to chance.

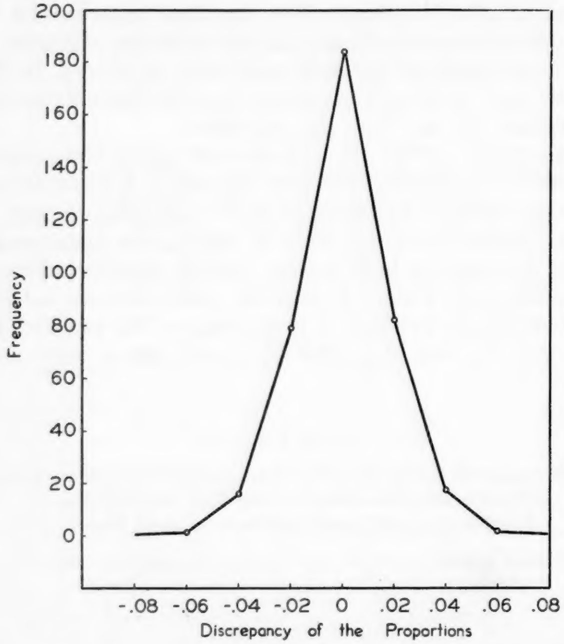


FIGURE 5
Distribution of the Discrepancies Between Actual and Theoretical Proportions

CONCLUSIONS

1. This experiment, using the paired comparison method, has proved that it is possible to scale individuals taking any mental and educational test. Thus, it is possible to utilize the psychophysical methods in the mental test situation.

2. In this particular problem it was found that, of three possible ways of dealing with the intermediate category, the method of using the proportion obtained by dividing this category equally between the other two was most satisfactory. It may be said that while the two-category judgments are preferred to three-category judgments in the ordinary psychophysical problem, it is impossible to use only two

categories here. We cannot avoid the fact that in the cases where two hypothetical individuals have the same number of successful persons they must be judged equal.

3. In applying Case V of the Law of Comparative Judgment to the data of this study, the two assumptions of normality of the distribution of the discriminal processes (variability of the individuals) and of equality of the discriminal dispersions have been actually subjected to test, and the assumptions have been found to be completely valid for these data.

4. The weighted scale and the unweighted scale were compared, and the difference between the mean differences was found to be insignificant.

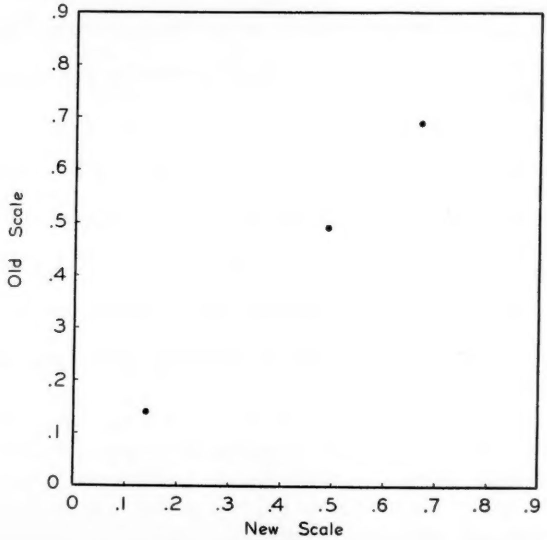


FIGURE 6
Distances Between Any Two Hypothetical Individuals in the Old Scale Plotted Against the Distances Between the Same Individuals in the New Scale

5. For the proof of internal consistency three tests were applied, one based on observational check on the columns of differences, one based on the algebraic average discrepancy and the average discrepancy, disregarding sign, and a test of stability. The results clearly indicate that the scale is internally consistent.

6. The test of stability here introduced may be said to confirm Professor Thurstone's method of internal consistency. So far as the writer knows, this type of test has not heretofore been attempted.

REFERENCES

1. Fisher, R. A. Statistical methods for research workers. 7th ed. London: Oliver & Boyd, 1938.
2. Guilford, J. P. Psychometric methods. New York: McGraw-Hill Book Company, 1936.
3. Lindquist, A. E. Statistical analysis in educational research. Boston: Houghton Mifflin Company, 1940.
4. Mosier, Charles I. Psychophysics and Mental Test Theory: Fundamental Postulates and Elementary Theorems, *Psychol. Review*, 1940, **47**, 355-66.
5. Peatman, John G. On the meaning of a test score in psychological measurement, *Amer. J. Orthopsychiatry*, 1939, **9**, 23-47.
6. Thurstone, L. L. A law of comparative judgment, *Psychol. Review*, 1927, **34**, 273-86.
7. Thurstone, L. L. A mental unit of measurement, *Psychol. Review*, 1927, **34**, 415-23.
8. Thurstone, L. L. An experimental study of nationality preferences. *J. Gen. Psychol.*, 1928, **1**, 405-25.
9. Thurstone, L. L. Attitudes can be measured, *Amer. J. Sociol.*, 1928, **33**, 529-54.
10. Thurstone, L. L. Equally often noticed differences, *J. Educ. Psychol.*, 1927, **18**, 289-93.
11. Thurstone, L. L. Psychophysical analysis, *Amer. J. Psychol.*, 1927, **38**, 368-89.
12. Thurstone, L. L. The measurement of opinion, *J. Abn. Soc. Psychol.*, 1928, **22**, 415-30.
13. Thurstone, L. L. Three psychophysical laws. *Psychol. Review*, 1927, **34**, 424-32.
14. Trealor, A. E. Elements of statistical reasoning. New York: John Wiley & Sons, 1939.

PSYCHOMETRIKA—VOL. 7, NO. 1
MARCH, 1942

A METHOD OF ESTIMATING ACCURACY OF TEST SCORING

WALTER L. DEEMER
HARVARD UNIVERSITY, CAMBRIDGE, MASSACHUSETTS

When errors of test scoring obey a Poisson frequency law (theoretical considerations suggest that they do), the method described may be used for finding the upper fiducial limits of scoring errors per paper. A criterion is suggested for establishing tolerance limits on scoring errors, and a method is given (1) for finding the probability of being wrong in the statement that the tolerance limit is being met for a given size sample or (2) for finding the size of sample that will make this probability not greater than some fixed value.

Most methods of scoring tests are liable to scoring error. Tests scored by test scoring machines are probably as little liable to error in scoring as any, but even such test scores may contain errors due to faint markings, or to stray dots on the answer sheet, or to misreading of the dials on the machine. The number of scoring errors in manually scored tests may be relatively large, and in a project of any size, where many tests are to be scored, it is often desirable that samples of scored tests be rescored in order to estimate the number of scoring errors that are being made. An estimate of scoring errors is necessary in evaluating later findings, and such an estimate may make it possible to make adjustments in the scoring methods if the sample indicates that too many errors are being made. This paper deals with the problem of estimating the number of scoring errors in a set of papers from the number of scoring errors found in a sample.

The method is based on the assumption that the number of scoring errors per paper follows the Poisson frequency law. From *a priori* considerations this seems a reasonable assumption for most tests, since the number of items is generally large and the probability of making a scoring error on any item is small and presumably constant from item to item. For any given test situation the observed frequency of scoring errors should be tested against the Poisson distribution, using a chi-squared test.

The number of opportunities for making errors in scoring may be different from the number of items in the test. If the mean number of scoring errors per paper is $\bar{x}$ and there are n items in the test, the assumption that the probability of making a scoring error is

$$q = \bar{x}/n \tag{1}$$

may not be valid. The best method of estimating q (the probability of making an error) and n (the number of opportunities for making scoring errors) is to compute the mean, $\bar{x}$, and the variance V, for the sample, and use the relations

$$q = 1 - \mathrm{V}/\bar{x} \tag{2}$$

$$n = \bar{x}^2/(\bar{x} - V) \ , \tag{3}$$

which are found by solving the formulas for the binomial,

$$\bar{x} = nq \tag{4}$$

$$V = nq(1 - q), \tag{5}$$

for n and q.

It is clear that the binomial estimated from (2) and (3) will, for $\bar{x} < V$, be of the form

$$[-q + (1 + q)]^{-n}, \tag{6}$$

which is called by Whitaker (9) and Pearson (4) a "negative binomial."

Whether the binomial estimated from (2) and (3) is close enough to the Poisson distribution to warrant the assumption that the population is distributed according to the Poisson law is discussed by Whitaker (9) and "Student" (7). Pearson (4) gives a method of resolving a series giving a negative binomial into a sum of two Poisson series. The rest of this paper is based on the assumption that the fit of the data in hand to a Poisson distribution has been found satisfactory.

Consider an example of the problems that arise in practice when we are trying to estimate the number of scoring errors that have been made in a set of papers. Let N_t be the total number of papers in the set, and N_s be the number of papers in the sample rescored. If x scoring errors are found in the N_s papers, what estimate, m_u, may be made of the upper limit of the mean number of scoring errors in the N_t papers, if a certain probability of being wrong in the estimate is acceptable?

The usual method of fiducial inference leads to two estimates of a parameter, an upper limit and a lower limit, such that in the long run of trials the statement that the true value of the parameter lies between these two limits will be wrong less than 100 p% of the time, where p, between 0 and 1, is called in this paper the level of significance ($1 - p$ is sometimes called the fiducial coefficient or the confidence coefficient).

Since we are interested here only in the upper fiducial limit, the problem is somewhat modified. The parameter we are interested in is the mean number of scoring errors per paper in the population. We shall denote this parameter by m. We wish to find a value of m, say m_u, such that any hypothesis that $m \geqslant m_u$ may be rejected at the chosen level of significance. Stated another way, we want to find a value of m, say m_u, such that the statement that $m < m_u$ will be right, in the long run, $100(1-p)\%$ of the time. The value of p is at the choice of the experimenter, but as p is made small m_u becomes large.

The problem of estimating the upper limit of the number of scoring errors in a set of papers is thus seen to be one of finding fiducial limits for the parameter, m, of the Poisson distribution. A number of papers have dealt with this problem. Ricker (6) and Garwood (2) consider the problem of finding upper and lower fiducial limits for the parameter of the Poisson distribution. Przyborowski and Wilenski (5) have considered the problems which arise in finding upper fiducial limits for the Poisson distribution.

The following theorem therefore proves nothing new, but it makes the problem of setting up fiducial limits for the Poisson somewhat easier to follow.

Theorem 1. If the number of scoring errors per paper is distributed according to a Poisson law with parameter m, then the sum of the number of scoring errors in a sample of size N_s will obey a Poisson law with parameter mN_s (see Uspensky (8), p. 279). This may be proved as follows. The Poisson distribution is

$$f(x) = m^x \frac{e^{-m}}{x!}. \tag{7}$$

The characteristic function of the Poisson is defined as

$$\phi(t) = e^{-m} \sum_{x=0}^{\infty} e^{itx} \frac{m^x}{x!} \tag{8}$$

$$= e^{-m(\cos t + i \sin t - 1)}. * \tag{9}$$

Since the characteristic function of a sum is the product of the characteristic functions, we get for the characteristic functions of the sum of a sample of size N_s:

$$e^{-N_s m(\cos t + i \sin t - 1)} \tag{10}$$

and, since "the distribution function of probability is uniquely determined by the characteristic function " (8, p. 271), the theorem is proved.

* The writer is indebted to Dr. L. Alaoglu for suggesting that to evaluate the sum in (8) let $z = me^{it}$.

This theorem gives us the information we need to find the upper fiducial limit of the parameter of the Poisson distribution. Let $\bar{x}$ be the mean number of scoring errors per paper found in the sample. The probability of getting a sample of N_s papers giving $N_s\,\bar{x}$ or fewer scoring errors from a population with parameter m_1 is, by Theorem 1,

$$p_0 = \sum_{x=0}^{N_s \bar{x}} \frac{e^{-m_1 N_s}(N_s\, m_1)^x}{x!}. \tag{11}$$

If $p_0 \leq p$, where p is the level of significance, the hypothesis that $m \geqslant m_1$ is rejected. If $p_0 = p$, we call m_1 the upper fiducial limit and denote it by m_u.

TABLE 1*

Upper fiducial limits, m', of scoring errors per sample of size N_s

$b \diagdown p$	0.50	0.10	0.05	0.02	0.01	0.005	0.001	$p \diagup b$
0	0.7	2.3	3.0	3.9	4.6	5.3	6.9	0
1	1.7	3.9	4.7	5.8	6.6	7.4	9.2	1
2	2.7	5.3	6.3	7.5	8.4	9.3	11.2	2
3	3.7	6.7	7.8	9.1	10.0	11.0	13.1	3
4	4.7	8.0	9.2	10.6	11.6	12.6	14.8	4
5	5.7	9.3	10.5	12.0	13.1	14.2	16.5	5
6	6.7	10.5	11.8	13.4	14.6	15.7	18.1	6
7	7.7	11.8	13.1	14.8	16.0	17.1	19.6	7
8	8.7	13.0	14.4	16.2	17.4	18.6	21.2	8
9	9.7	14.2	15.7	17.5	18.8	20.0	22.7	9
10	10.7	15.4	17.0	18.8	20.2	21.4	24.1	10
11	11.7	16.6	18.2	20.1	21.5	22.8	25.6	11
12	12.7	17.8	19.4	21.4	22.8	24.1	27.0	12
13	13.7	19.0	20.7	22.7	24.1	25.5	28.4	13
14	14.7	20.1	21.9	24.0	25.5	26.8	29.9	14
15	15.7	21.3	23.1	25.2	26.7	28.2	31.2	15
20	20.7	27.0	29.1	31.5	33.1	34.7	38.0	20
25	25.7	32.7	34.9	37.5	39.3	41.0	44.6	25
30	30.7	38.3	40.7	43.5	45.4	47.2	51.1	30
35	35.7	43.9	46.4	49.4	51.4	53.3	57.4	35
40	40.7	49.4	52.1	55.2	57.4	59.4	63.7	40
45	45.7	54.9	57.7	61.0	63.2	65.4	69.8	45
50	50.7	60.4	63.3	66.7	69.1	71.3	76.0	50

b = number of scoring errors found in a sample of size N_s.

p = level of significance; if the statement is made that the mean number of errors per paper, m, in the total set is less than m'/N_s, the statement will be true in the long run of trials, $100(1-p)\%$ of the time.

* Except for the column for $p = 0.50$, this table is condensed from a table given by J. Przyborowski and H. Wilenski (5, p. 288).

The expression (11) may be evaluated by use of the following relationship. Let $N_s m_1 = m'$ and $N_s x = b$; then

$$p_0 = \sum_{x=0}^{b} \frac{1}{x!} e^{-m'} (m')^x = \frac{1}{b!} \int_{m'}^{\infty} x^b e^{-x} \, dx , \tag{12}$$

as may be shown by integrating the right-hand member by parts. The value of the right-hand member may be found in *Tables of the Incomplete Gamma Function,* edited by Karl Pearson. Values of $m' (=N_s m_u)$ for 7 values of p and for values of $b (=N_s \bar{x})$ from 0 to 50 are given in Table 1, which has been condensed somewhat from 5, Table V. The column for $p = 0.50$ is not given in 5.

Example 1. Assume that in a sample of $N_s = 3$, one scoring error is found. Then $b = 1$. Say p has been chosen as 0.01; then Table 1 is entered with $p = 0.01$ and $b = 1$, and the tabled value is $m' = 6.6$. This means that the hypothesis that $N_s m \geqslant 6.6$ is rejected at the chosen level of significance, but that the hypothesis that $N_s m < 6.6$ may not be rejected. The upper fiducial limit of scoring errors per paper is therefore $m_u = 6.6/N_s = 2.2$. If we say that the mean of the Poisson distribution of which we have a random sample is less than m_u, we shall be wrong, in the long run, in not more than 1% of trials.

As N_s is increased, the upper confidence limit is decreased for a given $\bar{x}$. Thus, if $\bar{x}$ remains 1/3 but N_s is increased to 12, we enter Table 1 with $p = 0.01$ as before, but b is now 4. We find $m' = 11.6$, giving $m_u = 11.6/12 = 0.967$. This means that the hypothesis that $m \geqslant 0.967$ may be rejected at the 1% level, or, in other words, that the statement that $m < 0.967$ will be false, in the long run, not more than 1% of the time.

The factors to consider in deciding on size of sample will be the relative importance of having m_u accurately determined, compared to the cost of rescoring. The minimum value of N_s will be one that will give for $b = 0$ a value of $m_u \leqslant T$, where T is the maximum number of errors per paper considered acceptable. T will be called the tolerance limit of scoring errors per paper. This minimum value of N_s will be denoted by N'_s. It is found as follows. Enter Table 1 with $b = 0$ and p equal to whatever value has been chosen for the level of significance; the tabled value, m', is equal to $N'_s m_u$. Since we want $m_u \leq T$ we have

$$N'_s \geqslant m'/T . \tag{13}$$

Example 2. If T is 2, and $p = 0.01$ we find $m' = 4.6$ and $N'_s > 4.6/2 = 2.3$. The next largest integral value is taken when N'_s is not integral, and 3 papers would therefore be drawn at random from the set of N_t and rescored. If no errors in scoring were found in these three papers we would be confident at the $1 - p$ level that the tolerance limit T was not being exceeded.

Example 3. If one error is found during the scoring of the three papers taken in Example 2, m' changes. The new value of m' is found by entering Table 1 with the same p, but with b now equal to 1, since b is the number of scoring errors found in the N_s papers rescored; m' is found to be 6.6. Hence m_u is $6.6/3 = 2.2 > T$. There are now two courses open: (1) we may assume that the tolerance limit T is not being met and therefore change the method of scoring to make it more precise, or (2) we may take a larger sample in order to get a more precise estimate of m_u before making any inferences about T. If the scores of the group and not the individuals are the important thing, and if we may assume that no scoring errors remain in a paper after it has been rescored, we may score enough papers so that for the total group of N_t papers the upper estimate of the number of errors per paper is not greater than T. If we denote this upper limit for the whole set (after N_s have been rescored) by m_w, we have

$$m_w = m_u \left(\frac{N_t - N_s}{N_t} \right). \tag{14}$$

If we set $m_w \leq T$ and solve for N_s we get

$$N_s \geqslant \frac{m' N_t}{N_t T + m'}. \tag{15}$$

Hence we see that, if we are interested only in having $m_w \leq T$ without respect to the size of m_u, we may always rescore enough papers to satisfy .this requirement. But even in this case, m_u should be examined to see if it is so large as to indicate a possibility that more accurate scoring is possible.

Criteria for Choosing T. It seems clear that the value of T should depend to some extent on the variability of the scores in the group. If the variance of the scores is large, a slight error in a score due to scoring errors will not be so important as the same error when the variance of the scores is small. T may also be a function of the reliability of the test, as the more reliable the test the more important is a given change of score due to scoring errors. Professor T. L. Kelley has suggested to the writer that the most valid criterion of T will generally be some function of the standard error of a test score. If s is the standard deviation of the scores in the set and r is the reliability coefficient of the test, we have for the standard error of the test score

$$s_e = s\sqrt{1-r}\,. \qquad (16)$$

If the reliability coefficient is not known when the scoring is being checked, it may be estimated.

In line with a proposal about significant figures (3), Professor Kelley suggests that the most serviceable criterion would be one in which the median number of scoring errors was 1/3 s_e.

This amounts to choosing

$$T = 1/3\, s_e \qquad (17)$$

and using $p = 0.50$. If we choose a higher significance level, T would be increased. If the distribution of scoring errors were normal, the approximate equivalent of $p = 0.50$, $T = 1/3\, s_e$, would be $p = 0.05$, $T = 1.0\, s_e$, or $p = 0.01$, $T = 1.3\, s_e$.

Example 4. The approximate s_e for the Paragraph Meaning Test of the Stanford Achievement Test Advanced Battery is 3.7 items; $T = 1/3\, s_e = 1.2$. If one paper were chosen at random from a set and no scoring errors were found, we would be satisfied that the tolerance limit was being met, as Table 1 gives $m' = 0.7$ for $b = 0$, $p = 0.50$; hence $m_u = 0.7 < 1.2$.

Intuitively we may feel that 0 errors in a sample of 1 is not sufficient evidence that m, the mean number of errors in the total set of papers, is not greater than 1.2; it should be kept in mind that since we have used $p = 0.50$, our evidence is only that in half the trials $m < 1.2$. Reference to the column headed $p = 0.10$ in row $b = 0$ shows that in 10% of trials $m \geqslant 2.3$, and in 1% $m \geqslant 4.6$. Some workers may no doubt feel that a more rigorous criterion, say $T = 1/3\, s_e$ with $p = 0.05$ or $p = 0.01$ will more nearly fill their requirements in estimating scoring errors.

If $b (= N_s\bar{x}) > 50$, Table 1 may not be used. In this case only a slight error will result if it is assumed that $\bar{x}$ is distributed normally about mean m_u with variance m_u. (In the Poisson distribution, the mean equals the variance). The usual method of estimating upper fiducial limits for the normal curve gives

$$m_u = \bar{x} + z\sqrt{\frac{m_u}{N_s}}, \qquad (18)$$

where z is the distance in standard deviation units from the mean to the point cutting off a tail containing 100 p% of the area of a normal curve. Solving (18) for m_u we get

$$m_u = z\sqrt{\frac{\bar{x}}{N_s} + \frac{z^2}{4N^2_s}} + \left(\bar{x} + \frac{z^2}{2N_s}\right) \qquad (19)$$

or

$$m' = z\sqrt{\bar{x}\,N_s + \frac{z^2}{4}} + \left(\bar{x}\,N_s + \frac{z^2}{2}\right). \tag{20}$$

Table 2 gives a comparison of the values of m' from Table 1 and from the normal curve using (20) for $b(=N_s\bar{x}) = 50$.

TABLE 2

p:	0.50	0.10	0.05	0.02	0.01	0.005	0.001
m' from Table 1 $b = 50$	50.7	60.4	63.3	66.7	69.1	71.3	76.0
m' from (20), $b = 50$	50.0	59.9	63.1	66.8	69.4	71.8	77.1

It is seen that use of (20) for $b = 50$ leads to under-estimation of m' only when $p > 0.02$, and is only slightly under for $p = 0.05$.

The larger the set from which the sample is taken, the smaller the proportion of papers needed in the sample in order to satisfy (15). For example if $N_t = 10$, $b = 1$, $p = 0.05$ and $T = 0.48$, we find from Table 1, $m' = 4.7$. By (15) this gives

$$N_s = (4.7)(10)/[(10)(0.48) + 4.7]$$
$$= 4.9\,,$$

meaning that half the total number of papers would have to be rescored.

If all the foregoing figures remain the same except that N_t is 100, (15) gives:

$$N_s = (4.7)(100)/[(100)(0.48) + 4.7]$$
$$= 8.9\,,$$

and only 9 % of the papers need to be rescored. It is clear, therefore, that the scoring procedure should be so arranged that the probability of scoring errors will be constant for as large a number of papers as possible. In general, it will not be valid to combine papers scored by more than one scorer, as the probability of errors is likely to vary from one worker to another. If fatigue affects scoring accuracy, it may not be safe to assume that all papers from one scorer for a single long scoring period have the same probability of scoring errors. There are thus limits on the size of set which can be considered homogeneous with respect to the probability of scoring errors. Homogeneity may be tested by finding the binomial from the expressions (2) and (3) and seeing if q is small and n large. For a discussion of this problem, see Whitaker (9). If $q > 0.01$, the assumption that the population is distributed according to the Poisson law may not be sound. It may then be better to assume that the population has a binomial distribution. Upper fiducial limits for the binomial may be found from charts given by Clopper and Pearson (1, pp. 410 and 411).

An examination of the size of m_u may indicate that the methods of scoring are not sufficiently accurate, even though by using samples of size N_s based on (15) we are confident that the tolerance limit is being met for the set as a whole. In this case, economy of scoring may perhaps be secured by having the original scorers work more slowly, if that will increase accuracy, particularly if rescoring costs more per paper than the original scoring. This may be the case, for example, when the original scoring requires that the papers be marked.

The writer takes pleasure in expressing his gratitude to Professor T. L. Kelley for suggestions regarding some of the problems that arose in connection with this paper.

REFERENCES

1. Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934, **26**, 404-413.
2. Garwood, F. Fiducial limits for the Poisson distribution. *Biometrika*, 1936, **28**, 437-442.
3. Kelley, T. L. How many figures are significant? *Science*, 1924, **60**, 524.
4. Pearson, K. On certain types of compound frequency distributions in which the components can be individually described by binomial series. *Biometrika*, 1915, **11**, 139-144.
5. Przyborowski, J. and Wilenski, H. Statistical principles of routine work in testing clover seed for dodder. *Biometrika*, 1935, **27**, 273-292.
6. Ricker, W. E. The concept of confidence or fiducial limits applied to the Poisson frequency distribution. *J. Amer. Stat. Assoc.*, 1937, **32**, 349-356.
7. Student. An explanation of deviations from Poisson's law in practice. *Biometrika*, 1919, **12**, 211-215.
8. Uspensky, J. V. Introduction to mathematical probability. New York: McGraw-Hill Book Company, 1937.
9. Whitaker, L. On Poisson's law of small numbers. *Biometrika*, 1914, **10**, 36-71.